# Methodologies for Monitoring and Evaluating Vitamin A Deficiency Intervention Programs

A Report of the International Vitamin A Consultative Group (IVACG)®

The purpose of the International Vitamin A Consultative Group (IVACG) is to guide international activities aimed at reducing vitamin A deficiency in the world. The group offers consultation and guidance to various operating and donor agencies that are seeking to reduce vitamin A deficiency and its accompanying blindness. As part of this service, IVACG has prepared guidelines and recommendations for

- Assessing the regional distribution and magnitude of vitamin A deficiency
- Developing intervention strategies and methodologies to control vitamin A deficiency
- Evaluating the effectiveness of implemented programs on a continuing basis so that the evaluation of the effectiveness of intervention techniques is a continuing and dynamic procedure
- Research needed to support the assessment, intervention, and evaluation of programs

These guidelines and recommendations are available through IVACG's publications program. A list of publications available from IVACG, along with ordering information, is given on the inside back cover of this monograph.

The publication of this monograph is made possible through a cooperative agreement between the Office of Nutrition, Bureau of Science and Technology, Agency for International Development of the United States of America, and The Nutrition Foundation, Inc., Washington, D.C.

International Standard Book Number: 0-944398-04-9 July 1989

# Methodologies for Monitoring and Evaluating Vitamin A Deficiency Intervention Programs

## A Report of the International Vitamin A Consultative Group (IVACG)®


AUTHORS
Guillermo Arroyave, Ph.D.
Jane Baltazar, M.D.
Jane Kusin, M.D., Ph.D.
James M. Lepkowski, Ph.D.
Roy C. Milton, Ph.D.
S. G. Srikantia, D.Sc.

IVACG
Abraham Horwitz, M.D., M.P.H., Chairman
Frances R. Davidson, Ph.D., Secretary

IVACG Steering Committee
Barbara A. Underwood, Ph.D., Chairperson
Edward M. DeMaeyer, M.D.
Demissie Habte, M.D.
Florentino Solon, M.D., M.P.H.
Alfred Sommer, M.D., M.H.Sc.

IVACG Secretariat
C.O. Chichester, Ph.D.
Laurie Lindsay, R.D.


A

# Contents

I. Introduction ..... 1
A. Statement of the Purpose of the Manual
B. Basic Concepts
C. Types of Interventions

II. Components of the Evaluation Process ..... 5
A. Planning
B. Assessment of Operational Adequacy
C. Interpretation of Evaluation Data
D. Reporting

III. Description of Dietary and Biological Indicators for Evaluation ..... 11
A. Introduction
B. Suggested Indicators

IV. Evaluation of Specific Vitamin A Intervention Programs ..... 17
A. Nutrition Education
B. Horticulture Projects
C. Food Fortification
D. Periodic Distribution of a Large Dose of Vitamin A

V. Designs for Evaluation Studies and Sample Surveys ..... 29
A. Introduction
B. Study Designs
C. Sample Survey

Appendix ..... 37
Introduction
1. Sample Design
2. Sample Size Determination

Recommended Reading ..... 65

B

# I Introduction

## A. Statement of the Purpose of the Manual

The last few years have witnessed an increasing awareness of vitamin A deficiency as a public health problem. Epidemiological studies have emphasized its health implications for both individuals and populations. The geographic distribution of this nutritional disorder is extensive, affecting large sections of populations in developing countries, including Asia, Africa, and areas of Latin America.

Extensive knowledge has accumulated on the factors that condition and precipitate vitamin A deficiency. Inadequate consumption of animal foods rich in vitamin A or plant foods rich in carotenes is a central factor. In addition, factors have been identified that promote losses, increase physiological needs, and interfere with the intestinal absorption of the different vitamin A active compounds in the ingested food. These factors particularly affect the absorption of carotenes, which constitute the largest proportion of vitamin A in the diets of people in developing countries.

During the last two decades, attention has been paid to the development of strategies for the control and eradication of vitamin A deficiency as a public health problem. Various countries in the developing world where the problem has been identified have implemented intervention programs, and several other countries are planning to do so.

The published and unpublished records are plagued with examples of intervention programs that have been carried out without appropriate evaluation. This condition poses a serious limitation from both scientific and practical points of view. Most crucial among these limitations are: 1) it is not determined whether the intervention has produced the expected results for which it was planned, and 2) even when changes accompanying the intervention can be empirically visualized, it is not possible to estimate the extent to which those changes may be attributed to the program, e.g., observed changes could be caused by concomitant factors unrelated to the program. In addition, in the absence of proper evaluation, information will be lacking on the relationship between the magnitude of the total effort invested in the intervention and its degree of effectiveness.

Without evaluation, policy-makers lack the information necessary to decide whether a program to control vitamin A deficiency should be phased out, continued as is, or modified. Furthermore, the informed selection of alternative interventions to achieve the same objectives becomes impossible. Finally, the potential of the intervention for successful implementation in similar situations elsewhere (other countries or regions) would remain undetermined.

The need to present and promote evaluation as an integral component of any intervention, and to facilitate its implementation is clear. This manual is meant for those persons involved in intervention programs who are not experts in monitoring and evaluation. The purpose of evaluation in the context of this manual is to demonstrate effects or an absence of these from routinely implemented programs. Additionally, the evaluation results are important to the modification of programs in light of observed constraints. Rather than emphasizing the portion of knowledge still uncertain about a program, this manual emphasizes use of the knowledge available and its application for the evaluation of intervention programs to control vitamin A deficiency.

## B. Basic Concepts

Much has been written about evaluation in recent years. Different terminologies are often used to refer to the same concepts. An unfortunate consequence has been some confusion rather than clarification of the issues involved, especially among non-specialists. An effort has been made in this manual to avoid such confusion. A brief definition of the concept of evaluation and description of some basic terms used in this manual follows. Additional terms referring to the components of the evaluation process are presented in Chapter II.

1. Evaluation:
   Evaluation is a process by which the degree of success in achieving the predefined goals of an intervention is determined and the reasons for shortcomings in achievements identified. Thus, evaluation is an essential component to be built into any intervention plan.

2. Definition of terms:
   (Definitions and terminology used in this document conform, to the extent possible, with those used in Chapter III of *Guidelines for the Eradication of Vitamin A Deficiency and Xerophthalmia*, IVACG, 1977.)

   Activities and issues involved in an evaluation of intervention may be conveniently categorized as follows:

   *Inputs:* Actions which are taken to produce the desired changes.

*Outputs:* Intermediate changes seen as a result of the inputs and which ultimately lead to the outcome.

*Outcome:* The result of the intervention on the pre selected variable which it is expected to change according to the defined objective. An intervention program is expected to influence a chain of variables leading to an improved vitamin A nutritional status or a decrease in xerophthalmia (eye signs due to vitamin A deficiency). Whether the ultimate change or any of the intermediate changes is itself considered the outcome may be decided by the evaluation-designers on the basis of the primary objective of the program and of feasibility considerations, as explained further in Chapter III.

## C. Types of Interventions

To date there are two categories of intervention programs specifically addressing hypovitaminosis A and xerophthalmia. These are:

1. Interventions that increase the dietary intake of vitamin A by a population through:
   a. Nutrition education,
   b. Horticulture at the macro- and micro-level, and
   c. Fortification of foods in existing diets.

2. Interventions aimed at treating and controlling the prevalence of ocular damage due to vitamin A deficiency through periodic use of large dose supplements.

Apart from these, primary health care approaches to control concurrent disease factors which result in elevated vitamin A requirements and/or which interfere with the absorption and utilization of dietary vitamin A, also help in improving vitamin A status. In addition, any improvement in the socioeconomic status of the target population is likely to indirectly influence the overall health and nutritional status of that population, and by inference would result in improved vitamin A status.

# II Components of the Evaluation Process

The evaluation process consists of several components that have been grouped in this report under four broad headings: A) planning, B) assessment of operational adequacy, C) interpretation of evaluation data, and D) reporting. Outlined under each heading are practical steps that are intended as guidelines.

## A. Planning

The following is a description of the process to be considered in preparing an intervention program evaluation.

1. Definition of program objectives:
   As mentioned earlier, evaluation should be an integral part of any intervention program. The aims and objectives of the program form the basis for evaluation and therefore must be specifically and clearly stated. It is, however, possible to redefine the objectives during the course of the evaluation in order to limit it to one or more outputs of the intervention.

2. Selection of indicators:
   Indicators are variables that help to measure change. Such changes may represent the final program outcome or some intermediate events occurring between the application of the intervention and the final outcome. Indicators selected should be appropriate for the type of intervention and objectives.

3. Population target, criteria, and timeframe:
   Criteria must be established and program targets defined in order to assess the progress toward achieving the given objectives. This involves:
   a. Specification of the target population for the intervention program (e.g., preschool children, high-risk children, urban/rural residents),
   b. Specification of the amount of change (increase or decrease)

desired for a given indicator that the program is intended to achieve, and

c. Specification of the timeframe for achievement. The time periods suggested are somewhat arbitrary but they have been found to be helpful in the process of policy implementation—short-term (0.5–2 years), medium-term (3–5 years), and long-term (6 or more years).

*Example:*

Program: Food fortification with vitamin A.
Objective: To improve the vitamin A status of the population.
Indicator: Serum vitamin A level.
Criterion: Serum vitamin A level below 10$\mu$g/dl in no more than 5% of the high-risk population.
Timeframe: 2 years.

4. Study design:
Study design for evaluation of an intervention program includes both a statistical and a sampling design. The statistical design will usually be either a "before-and-after" intervention comparison, or a comparison of intervention areas with concurrent non-intervention areas. Data for these designs are obtained from sample surveys, for which the sampling design describes how representative data are to be collected and analyzed. Effects of competing explanations may be controlled in the design by avoiding known confounding factors and by statistical adjustment in analysis. These and other issues in statistical design are discussed in Chapter V of this manual. In addition, a detailed discussion of sample design and statistical implications and of sample size determination is available in the Appendix.

5. Data type, source, and use:
Evaluation is based on valid, accurate, and reliable data. Validity means that the data actually measure what they are supposed to measure. Reliability implies that measurement of an indicator by different people at different times and under different circumstances yields essentially similar results. Accurate measurements are close to true values. Inaccuracies may occur due to imprecise instruments or to random variations in measurement techniques.

a. Type of indicator data:
Indicator data may be clinical (e.g., Bitot's spots), biochemical (e.g., serum retinol levels), dietary (e.g., frequency of consumption of carotene-rich foods), and logistical (e.g., resources). All such data should be appropriately documented.

b. Sources:
The data obtained from routinely collected statistics (e.g., information from growth charts) in the health infrastructure system may be used if they are considered reliable. If not, special studies or surveys may need to be conducted.

c. Use:
Data may be used for the following purposes:
(1) To provide baseline information (e.g., pre-intervention prevalence of Bitot's spots),
(2) To assess input (e.g., number of vitamin A capsules available at the health post),
(3) To permit monitoring (e.g., periodic recording of distribution of vitamin A capsules),
(4) To assess output (e.g., number of children receiving vitamin A capsules), and
(5) To determine outcome (e.g., post-intervention prevalence of Bitot's spots).

## B. Assessment of Operational Adequacy

Program evaluation comprises 1) assessment of adequacy of implementation, i.e., the type and number of activities, the personnel involved and the way the activities were carried out, and 2) assessment of outcome, i.e., the effect or impact of the program.

Assessment of the first component is important to improve the program design and the delivery of services. Assessment of the second is essential to convince policy-makers that the program was properly selected and did, in fact, achieve the objectives at least in part. The two components are often assessed simultaneously, although it would be desirable to conduct outcome evaluation after monitoring has shown that the program is well implemented.

1. Implementation:
The adequacy of implementation should take into consideration the following:
a. Inputs such as materials, personnel, and resources.
*Example:* For a six-month vitamin A capsule-distribution program for villages: Were there sufficient capsules in stock? Were the vehicles in running condition? Was there sufficient petrol available for each delivery schedule? Was the staff to distribute the capsules sufficient in number and adequately trained?
b. Activities such as adherence to planned schedules.
c. Outputs such as the availability and utilization of services by the target population.
*Example:* Vitamin A capsules may be available at the health center but this facility is too distant to serve a large section of

the target population. For this and other reasons, coverage may be small and below the critical level needed for a positive response.

d. Training and supervision:
   (1) Type and number of personnel trained, methodology and adequacy of training.
   (2) Organization of refresher courses.
   (3) Adequacy of supervisory structure at different levels.

e. Monitoring: Type of data collected and frequency of collection.

Awareness by functionaries at all levels about the details of the intervention and the philosophy behind it as well as awareness of the community concerning the program are important determinants of outcome, particularly in some types of intervention. Assessment of this "awareness status" should be considered a part of adequacy of implementation. Adequacy of implementation is judged against the pre-established levels of inputs and outputs met.

2. Outcome:
   Attainment of program objectives and targets is assessed according to evaluation designs discussed in Chapter V. Proper attention should be given to the influences of confounding factors on outcome.

## C. Interpretation of Evaluation Data

In the evaluation of some interventions more than a single type of indicator may be used. For instance, to assess the impact of a food fortification program aimed at improving the vitamin A nutritional status of the population, the evaluators—if resources are not a limiting factor—may decide to utilize dietary and clinical indicators in addition to the primary indicator suggested in the example given elsewhere (serum vitamin A levels).

If the integrated analysis and interpretation lead to changes in the same direction for all three indicators, this concordance synergistically enhances the validity of the conclusions. Such a procedure is desirable both from a scientific point of view and because of the level of confidence that the evaluation data generate. For example, when baseline serum retinol levels are found to be markedly low, the dietary intake of vitamin A is also inadequate and a high prevalence of clinical manifestations is seen. If, following the intervention, there is a reversal in all three indicators, the conclusion that the intervention had, in fact, been effective becomes very strong.

An absence of parallelism between the three types of data does not necessarily imply lack of relationship. Although the different parameters should reflect the existence of hypovitaminosis A in the population, the timing of the appearance of their deviation from normal is

often not concurrent at the individual level. In addition, mild and moderate degrees of deficiencies, as indicated by dietary and biochemical indicators, may be insufficient to induce clinical abnormalities.

Such a reinforcing multiple-indicator approach requires additional resources, and under the circumstances in many developing countries it will prove difficult to implement. Therefore it must be stressed, that although desirable it is not essential because the use of the minimum indicator chosen will generate the primary information required. Consequently, the lack of resources for attempting the multi-indicator approach should be no reason for discouraging the implementation of an evaluation strategy.

## D. Reporting

The last stage of the evaluation process is the preparation of the reports. For the scientific community, a detailed report with descriptions of every phase and activity is mandatory. This should include a brief description of the intervention being evaluated, the design and plan of the built-in evaluation, methodology and procedures for measurement, standardization and quality control, results, interpretations, and conclusions. This type of report should reflect the degree of scientific and technical rigor employed in such a manner that professionals reading it can judge the quality of the operation and establish the scientific soundness and reliability of the conclusions.

The judgement of the value of the intervention is essentially a professional opinion, and it is recommended that an independent panel of peers should review the total project. This procedure will not only add credence to the report, but it will also help to put in proper perspective any aberrant conclusions that may have been drawn by the evaluators.

For those who are not specialists but who will be in charge of making decisions on the study's recommendations, such as public health authorities and administrators, another type of report is necessary. The emphasis here should be on the results obtained and their significance in terms of the original objectives of the vitamin A intervention. With satisfactory monitoring and initiation of timely corrective steps, shortfalls in outcome should be minimal, but if for any reason monitoring was inadequate and outcome disappointing, the report should focus attention on the causes for the shortfall in outcome. This will help provide better implementation of similar, subsequent programs. Comments on the relative merits and disadvantages of the intervention studied, particularly in comparison with other approaches attempting the same objectives, should be included wherever such data are available. The specialist must present information in terms understandable to the non-specialist.

In the report directed to policy-makers and other interested

parties, it is prudent to include an executive summary, either as a separate document or preferably as the first section of the evaluation report. This should contain brief descriptions of the program being evaluated, of the methods used for evaluation, and of the findings of the evaluation study. For emphasis, the summary of the findings should be presented first. The body of the evaluation report can be effectively presented by following some variant of the following outline:

I. Background of the Program
   A. Origin of the program covering such items as enabling legislation, funding sources, and major personnel associated with initiating the program.
   B. A brief chronology of the program, covering dates of inception, major events, and origins of the evaluation.

II. Goals of the Program

III. Target Population

IV. Characteristics of the Program
   Description of the program, resources available, and the services that are to be delivered.

V. Description of the Evaluation Study, which includes:
   A. Purposes of the Evaluation
   B. Evaluation Design Used
   C. Outcome Measures Used
   D. Data Collection Procedures

VI. Evaluation of Results

VII. Discussion of Results
   A. Evaluation of Findings
   B. Generalizability of Evaluation Findings

VIII. Conclusions and Recommendations

# III Description of Dietary and Biological Indicators for Evaluation

## A. Introduction

Indicators are characteristics of the population that the intervention intends to change. This means that the indicators selected must have been shown by previous scientific studies to be connected with the intervention inputs in a cause-effect relationship.

Basic to the selection of appropriate indicators is a good understanding of the sequence of events that may be considered effects of the program inputs to control vitamin A deficiency (i.e., dietary, biochemical, clinical, and demographic). This sequence may be conceived as follows:

1. An increase in the intake of pre-formed vitamin A (retinol), or carotenes, or both;

2. An increase in vitamin A levels in body fluids and tissues; and

3. The correction of physiological alterations, including reversal of tissue and organ damage (clinical) known to be due primarily to vitamin A deficiency.

Growth retardation, loss of appetite, and increased morbidity and mortality have also been described as manifestations of vitamin A deficiency and have therefore been suggested as indicators. None of these however is specific to vitamin A deficiency, and the usefulness of measuring changes in these indicators as evidence of altered vitamin A status is very uncertain.

## B. Suggested Indicators

The selection of indicators for evaluation of interventions must be based not only on scientific knowledge but also on what is perceived to be feasible for implementation at the time. An excessively ambitious intervention goal and evaluation plan may overwhelm the existing

administrative and technical capacity available in the developing countries. This would possibly result in discouraging and thereby delaying decision-making in the implementation of potentially good programs. A brief discussion of indicators that can be used is presented below.

(Methods and procedures for the laboratory measurement of vitamin A in these biological materials have been published in detail by IVACG in *Biochemical Methodology for the Assessment of Vitamin A Status.*)

1. Dietary Intake:
   Dietary intakes of preformed vitamin A and carotenoids are not indicators of vitamin A status in the same sense that the biochemical and clinical indicators are. Hypovitaminosis A and xerophthalmia are, however, more prevalent in population groups with inadequate intake of vitamin A-containing foods.

   Data relevant to habitual intakes of vitamin A by the population (items a and b below) and more specifically by the infant and young child (c and d) can be obtained by the following approaches:

   a. Assessing the frequency of consumption of carotene-rich foods and/or preformed vitamin A by the target population,
   b. Calculating intake of vitamin A (micrograms or retinol equivalents per day) by the target population,
   c. Determining the age at which and the relative amount of additional foods containing carotene and/or preformed vitamin A are first given to children aged 0-3 years, and
   d. Determining the percentage of children aged 0-3 years breast fed by age groups (3 month intervals).

   In addition to being an indirect indicator of vitamin A status, use of dietary information is limited by consideration of accuracy and reliability of estimates of habitual intakes. Factors important in this context are:

   a. Errors of varying magnitude inherent in practically all types of dietary assessment methods,
   b. Marked seasonal fluctuations in the intake of carotenoid-rich foods that make it difficult to interpret data obtained from cross-sectional (single point) surveys,
   c. The limited information available on provitamin A content of foods as eaten and variability of concentration of vitamin A in breast milk,
   d. The limited information on bioavailability of carotene from habitual diets, and
   e. The probable effect of culinary practices on vitamin A content of diets.

For purposes of monitoring and evaluation of an intervention aimed at increasing the consumption of vitamin A and/or carotene-rich foods, these limitations are not a major concern provided that dietary information is collected in a standardized manner. More important than the actual amounts are alterations in the consumption of these foods that can be used as a measure of outcome.

Even in intervention programs where more objective indicators are used, such as in periodic large-dose vitamin A distribution programs and in fortification of food items with vitamin A, dietary information on intakes of the vitamin should be included 1) to monitor trends in dietary intake over time to determine if they influence the vitamin A status of the recipients and therefore affect outcome of the intervention, and 2) to assess when the programs should be modified.

Collection of dietary information is very important. Although the 24-hour recall method and the 3- or 7-day diet history of estimating dietary intake of vitamin A are often considered easy to perform, they are not recommended for purposes of evaluation for population-based studies. Though it is traditionally recommended that diet surveys not be done during sickness, it is important to get a more realistic estimate of habitual dietary intake. Therefore, diet surveys should be done irrespective of the health status of the child.

Because in rural populations agricultural cycles can alter dietary intakes, recordings should be done periodically, and frequency of data collection should take into account this factor. If resources are limited, recordings can focus on specific seasons of anticipated shortages of foods containing vitamin A and/or peak disease incidence.

In urban populations, the timing of data collection will be governed by other considerations including the effect of working days versus weekends, and before and after salary is received.

2. Biochemical:
   Biochemical indicators often reflect vitamin A nutritional status before clinical abnormalities can be detected. Reduced levels of vitamin A in body fluids and tissues are usually seen before metabolic functions and tissue damage occur. (See IVACG manual noted above for discussion of biochemical methodology for assessing vitamin A status.)

   Serum vitamin A: The most valid indicator of vitamin A nutritional status is the store of the vitamin in the liver. Because this is not

easy to determine, serum vitamin A level is used as an alternate indicator. Levels below 20 μg/dl are frequently associated with unsatisfactory liver reserves, whereas levels below 10 μg/dl are indicative of very low reserves. A shift to the right in the lower portion of the frequency distribution of serum vitamin A levels in a population is therefore indicative of changes in vitamin A and can thus be used as a measure of the extent to which an intervention program has succeeded. A practical problem in some of the countries is the possible resistance encountered in obtaining samples of blood for analysis. Because both vitamin A and retinol binding protein (RBP) estimations can be carried out on finger-prick blood samples, this problem may not be as serious as it initially appears. Efforts should be made through persuasion to get the cooperation of the community.

Breast milk vitamin A: An additional biochemical indicator that can be used for intervention strategies targeted for populations which include women is the vitamin A concentration in breast milk. This method has the advantage of being non-invasive, and milk samples are easily obtained. Such collections should be made following a standardized procedure because of the diurnal and intrafeed variability in vitamin A content.

Unlike serum vitamin A, which has been widely used, breast milk retinol has been used infrequently. The efficacy of its use as an indicator in evaluation was demonstrated in the sugar fortification program in Guatemala. In this study, the criterion was the prevalence of breast milk samples with less than 20 μg/dl of vitamin A. Breast milk samples obtained from well-nourished women contain about 50 μg vitamin A per deciliter and the cut-off level of 20 μg/dl used in this study thus represented only 40% of the normal level. This cut-off point proved sensitive enough to reflect the impact of the intervention on lactating women. It is recommended that this criterion be followed until new knowledge indicates the need for the adoption of a different one.

Liver vitamin A: Direct measurements of levels of vitamin A in liver for assessment purposes are only possible in autopsy material. The use of this indicator is feasible only when there are no constraints in obtaining autopsy material and when the intervention strategy covers the entire population. Even then it is unlikely that the results will be truly representative of the population, though they may provide data on changing trends.

Recently a new method for measuring liver retinol levels indirectly—the relative dose response—has been developed. Its applicability under field conditions has been successfully demonstrated in a limited number of applications. Currently a

modification using a derivative of vitamin A (vitamin $A_2$) is being tested. This modification would require one rather than two blood specimens. If successful, this approach would be an important addition to the existing list of indicators.

3. Clinical:
   For a more detailed discussion see *The Symptoms and Signs of Vitamin A Deficiency and Their Relationship to Applied Nutrition.*

   There are five well-recognized clinical manifestations of vitamin A deficiency. These are all ocular—night blindness, conjunctival xerosis, Bitot's spots, corneal xerosis, and keratomalacia. In the absence of a history of trauma corneal scars are highly suggestive of previous vitamin A deficiency. Altered conjunctival histology as diagnosed by impression cytology is currently undergoing field tests to determine its suitability as a reliable and sensitive indicator of early vitamin A deficiency. Changes in the prevalence of one or more of these manifestations following an intervention may be used to evaluate the program outcome. (See IVACG manual noted above.)

   Ease and reliability of interpretation as well as prevalence rates of these manifestations have to be considered in the selection of the clinical indicator(s).

   The diagnoses of night blindness and conjunctival xerosis are prone to serious subjective errors, while that of Bitot's spots and corneal lesions are much less so. The very low prevalence rates of corneal lesions somewhat restricts their utility as an indicator, in contrast to that of Bitot's spots, which is almost always several fold higher. Unlike older children, in most preschool children Bitot's spots have been shown to respond to vitamin A. There are, however, instances of countries where a significant proportion of Bitot's spots even in young children do not respond to vitamin A administration and are thus not reliable indicators of vitamin A deficiency. It is therefore difficult to recommend a particular clinical indicator for universal use under all circumstances. In situations where the prevalence of Bitot's spots is high, it is the indicator of choice because of the ease of diagnosis. Though the presence of Bitot's spots neither impairs vision nor signifies impending corneal involvement leading to blindness, its selection can be justified on the grounds that earlier well-controlled intervention studies have shown that a reduction in the prevalence of Bitot's spots ran somewhat parallel to that of corneal lesions. In situations where the prevalence of Bitot's spots is low, but that of night blindness is higher, night blindness may be the preferred indicator despite the potential for subjective errors in diagnosis.

In such a case, all attempts must be made to reduce diagnosis errors.

Among vitamin A-deficient communities in several countries the prevalence of Bitot's spots among children has been found to increase significantly as age advances. Because the slope of this relationship changes as the vitamin A status improves, this can be used to roughly determine the effectiveness of an intervention program even in the absence of baseline data on the prevalence of Bitot's spots.

Hospital-based statistics collected over a period of time, in spite of their limitations, can sometimes be used to complement other evidences of effectiveness of vitamin A intervention programs. Change in the prevalence rates of ocular signs (particularly corneal lesions) in children who are seen in ophthalmic and pediatric hospitals may be used for this purpose.

# IV Evaluation of Specific Vitamin A Intervention Programs

This section covers the evaluation components of four types of intervention programs: nutrition education, horticulture projects, food fortification (nutrification), and periodic distribution of a large dose of vitamin A. For each intervention program, examples are given for the definition, objective, suggested indicators, criteria and targets, evaluation design, and a description of types and sources of data.

## A. Nutrition Education

1. Definition:
   Educational activity component of any project that encourages individuals, households, or schools to use their resources in a manner beneficial to improve vitamin A nutritional status.

2. Objective:
   To increase the consumption of vitamin A-containing foods. (Vitamin A deficiency usually exists as part of a multiple nutrient deficiency situation and it is necessary to realize that nutrition education has to aim at improving intakes of all nutrients.)

3. Indicators:
   a. Frequency of consumption per week of vitamin A-containing foods and/or carotene-rich foods by specific groups, i.e., the proportion of pregnant and lactating women and children aged 6–36 months eating foods containing vitamin A at least 3 times per week, and the proportion of infants eating vitamin A foods by the age of 12 months.
   b. Calculated intake of retinol equivalents per day, i.e., the proportion of pregnant and lactating women and also children aged 6-36 months having intakes adequate to meet recommended dietary allowances (RDA), and those having intakes which are at least 50% of RDA.
   c. Proportion of infants and children breast fed from birth to 18 months, with intervals of 3 months.

4. Criteria and targets:
   a. Duration of breast feeding until age 18 months for at least 75% of children, when possible.
   b. Introduction of additional foods containing vitamin A (e.g., dark green leafy vegetables or DGLV) at age 6-9 months for at least 75% of the age group, three times a week.
   c. Vitamin A intake from breast milk plus other foods reaching at least 50% RDA in all preschool children.

   There are limited data available to estimate whether these criteria can be reached by nutrition education per se and, if they can be achieved, the time needed. Under the guidance of motivated workers, successes have been reported within 5 years.

5. Evaluation design:
   a. Choice of design:
      The design of choice is the before and after comparison of the indicators of outcome. When reliable baseline data cannot be obtained despite all efforts, a comparison of non-project and project areas may be considered. (For details, see Chapter V and Appendix.)
   b. Adequacy of implementation:
      Input data collected should include the following:
      (1) The number of communicators trained, content and methodology of training, and back-up system.
      (2) The number of supervisors trained and efficiency of supervision.
      (3) The number of nutrition-education activities planned and executed, e.g., group discussions, demonstrations, and personal contacts.
      (4) Coverage.
      (5) Time exposure, i.e., duration and frequency of contacts between communicator and target groups. These data should be collected once every 3–4 months for the duration of the study.
   c. Adequacy of output:
      Data collection should include the following:
      (1) Household information:
         —to distinguish the effect of socioeconomic and cultural characteristics on response;
         —extended or nuclear family, female- or male-headed, education of those in the household, and occupations of husband and wife;
         —in a survey of randomly selected households from the target population, note participation in program by percent and compliance.

(2) Availability of vitamin A and/or carotene-rich foods in the local markets by calendar month.
(3) Use of vitamin A and/or carotene-rich foods by the household by calendar month.
(4) Duration of breast feeding of index child.
(5) Age at which additional foods are given to infants.
(6) Age at which vitamin A and/or carotene-rich foods are given to preschool children.
(7) Usual frequency of intake for preschool children and mothers (semi-quantitative).
(8) Knowledge and attitude regarding vitamin A and/or carotene-rich foods for preschool children and mothers, collected through specially constructed questionnaires.

6. Types and sources of data:
   a. Baseline data:
      Data on dietary intakes of women and feeding practices of infants and children may be obtained from mother and child health service records or from previous household surveys. Otherwise, they have to be collected from specially conducted surveys.
   b. Input data:
      These can be obtained from program directors and records of educators and supervisors. A sample survey in which staff performance and type/quality of activities are assessed and which includes cross-checking of a subsample of mothers will provide additional data.
   c. Project outcome:
      Repeat survey.

## B. Horticulture Projects

1. Definition:
   Programs that encourage and support the production and consumption of carotene-rich foods at the micro level (house, school, and community gardens).

2. Objectives:
   a. To improve the availability of carotene-rich foods at the family level, given the agro-ecological conditions of the area.
   b. To improve the consumption of carotene-rich foods by the family, particularly pregnant and lactating women and preschool children.

3. Indicators:
   Although the main objective of horticultural programs is to increase the availability of carotene-rich foods, all of them should also have a built-in education component to ensure the consumption of the garden produce. Some of the indicators used in evaluating horticultural programs, therefore, will be similar to those used in evaluating nutrition-education programs, described earlier. In addition, they will have other indicators that measure the availability of the carotene-rich foods.
   a. Number of home, school, and/or community gardens in existence.
   b. Production of specified carotene-rich foods, including the number of months of availability per year.
   c. Proportion of households eating carotene-rich foods grown at home, school, and/or community gardens.
   d. Frequency of consumption per week of vitamin A-containing foods and/or carotene-rich foods by age groups, i.e., proportion of pregnant and lactating women and children age 6–36 months eating carotene-rich foods at least 3 times per week; proportion of infants eating carotene-rich foods by the age of 12 months.
   e. Or, calculated intake of vitamin A per day, i.e., proportion of pregnant and lactating women and children aged 6-36 months having intake levels adequate to meet RDA and those having levels of at least 50% RDA.
   f. Proportion of infants and children breast fed from birth to 18 months with class intervals of 3 months.

4. Criteria and targets:
   a. Production of carotene-rich foods for at least 6 months of the year in semi-arid areas, and throughout the year if there are no constraints, by 95% of home gardens. For school and community gardens the percentage will depend upon the project objectives.
   b. Consumption at least 3 times per week, when carotene-rich foods are available, by at least 75% of the target groups concerned.
   c. Daily intake of vitamin A by at least 75% of women reaching at least 50% RDA.
   d. Daily intake by at least 75% of preschool children aged 6-36 months, reaching at least 50% of RDA.

   In well-implemented projects it should be possible to achieve an increase in the availability of carotene-rich foods in a period of two years.

Whether increased consumption of carotene-rich foods by preschool children can be achieved, despite increased availability, within this timeframe is uncertain. It would depend largely on the effectiveness of nutrition education in bringing about changes in traditional child-feeding practices. It is difficult, therefore, to suggest a timeframe for this component.

5. Evaluation design:
   Causal links between project inputs and outcomes are best assessed by a randomized trial that is seldom possible to conduct in this type of program. The design of choice is the before-and-after comparison of the indicators of outcome. When reliable baseline data cannot be obtained, a comparison of project and non-project areas may be considered. (For details, see Chapter V and Appendix.)

6. Types and sources of data:
   a. Baseline data:
      Data on infant- and child-feeding practices and dietary intakes of women may be obtained from mother and child health service records or from previous household surveys. Data on the number and type of gardens may be obtained from agricultural services in the program area. If not available from these sources, a specially conducted survey becomes necessary. (See IVACG's publication on dietary assessment guidelines.)
   b. Adequacy of output:
      Data on number of gardens started and type and amount of carotene-rich foods produced can be obtained from the records of the functionaries. This should be cross-checked, when possible, by actual observation in a subsample.
   c. Adequacy of outcome:
      Repeat survey.

## C. Food Fortification (Nutrification)

1. Definition:
   Programs consisting of the addition of vitamin A to conventional foods (usually processed) to increase the consumption of this nutrient by population groups.

2. Objectives:
   The objectives of fortification are:
   a. To raise vitamin A intake as close as possible to the level of recommended allowances.
   b. To improve the vitamin A status of the population.

3. Indicators:
   a. The primary indicator: concentration of serum retinol or RBP.
   b. Additional desirable indicators: breast milk retinol levels (from lactating women in the sample), liver retinol levels in autopsy material from selected morgues, and/or relative dose response among a subsample.

4. Criteria and targets:
   The critically low levels and their prevalence are as follows:

## Table 1

| Serum[a] | | Breast milk[b] | | Liver[c] | |
|---|---|---|---|---|---|
| μg/dl | Prevalence (%) | μg/dl | Prevalence (%) | μg/g | Prevalence (%) |
| <10 (Deficient) | 5 | n.k.[d] | n.k. | <5 | 3 |
| <20 (Low) | 15 | <20 | 15 | <20 | 15 |

[a]General population.
[b]Lactating women.
[c]Preschoolers.
[d]n.k. = not known.

A reduction in the prevalence of deficient levels in the population to levels well below the critical percentages given as "Criteria," in a time period of one to two years.

5. Evaluation design:
   a. Choice of design:
      Because vitamin A fortification programs are usually planned to cover a whole country or region, the use of a "control (non-intervention) versus intervention" design is usually not feasible. The design must necessarily be a "before-and-after" comparison of the indicators of outcome. As emphasized elsewhere (Chapter V), in this type of design it is essential to look for possible confounding factors. The most common are uncontrolled changes in vitamin A intake from natural foods or other sources, and marked increases in morbidity (e.g., measles and diarrhea), which can increase vitamin A needs and interfere with its absorption and utilization.
   b. Adequacy of implementation:
      (1) Input:
         Efficiency of fortification process at the fortification plants.

Extent to which quality standards are met is determined by:

—Control of uniform level of addition of the vitamin;

—Control of appropriate packing, storage, and handling of fortified product.

(2) Delivery system:

Continuous availability of the fortified food at distribution and retail outlets (markets, stores, etc.) all over the region where the target population lives.

(3) Output:

The output is the intake of vitamin A contributed by the fortified product. Assessing the adequacy of output involves therefore, the quantitative estimation of its consumption, particularly by the most vulnerable groups (e.g., preschool-aged children and pregnant and lactating women). Monitoring of this should be periodic at predetermined intervals. Extent of coverage is also calculated.

c. Adequacy of outcome:

Comparison of the indicators of outcome as measured at the end of the evaluation period with their baseline measurements, to determine the extent to which the criteria have been met in the specified timeframe.

6. Types and sources of data:

a. Baseline data:

(1) Concentrations of vitamin A in serum, breast milk, and liver; and prevalence of values below the critical levels. (Source: survey of the target population.) Liver vitamin A concentrations can be measured in autopsy material available from selected hospital morgues.

(2) Quantitative data on the consumption of the food or dietary item to be used as the vehicle for the fortification (e.g., sugar, skimmed milk powder, monosodium glutamate). (Source: specific consumption survey.)

b. Input data:

Material needs:

(1) Amounts of the vitamin A compound required. (Source of data: Earlier research on the level of fortification needed and the estimated total consumption of the food vehicle.)

(2) Number and location of fortification centers. (Source: Government or private records. The fortification centers are usually the factories where the food or dietary item is produced or processed.)

Manpower:

(1) Type and number of trained workers to carry out the fortification process.

(2) Type and number of workers needed for supervision and control of all phases and for keeping the corresponding records.

(3) Type and number of field and laboratory technicians for the collection and analysis of fortified food samples and specimens of biological materials for the determination of their vitamin A levels. (Source: Nutritional biochemistry laboratory that is serving the program.)

c. Data on monitoring:

(1) Continuous records of the amounts of the fortified food produced and distributed to the points of access for the populations, such as markets and stores. (Source: Government and/or private manufacturers' marketing records.)

(2) Continuous recording of "extraneous" factors that could be influencing baseline indicator values independently of the programmed input (e.g., vitamin A supplements, outbreaks of an epidemic, unexpected seasonal effects on availability of dietary sources of vitamin A). (Source: Direct observation by evaluation field team, records from health sector and other pertinent agencies.)

d. Data on output:

(1) Consumption of the fortified food or dietary vehicle by the target population. (Source: Specific periodic surveys. Because the actual vitamin A fortification level is known from analysis, the intake of additional vitamin A, above that eaten through habitual diets, can be calculated.)

e. Data on outcome:

(1) Post-intervention data on serum and breast milk vitamin A and prevalence of values below the critical levels. (Source: Follow-up survey of the target population; data on liver vitamin A to be collected from autopsy material obtained from the same morgues as in baseline.)

## D. Periodic Distribution of a Large Dose of Vitamin A

1. Definition:
   Projects in which large doses of vitamin A are periodically given to children below the age of five years.

2. Objective:
   a. To reduce severe ocular manifestations of vitamin A deficiency and to prevent their occurrence.

3. Indicators:
   The most suitable from among the following may be chosen:
   a. Bitot's spots—except in such areas where there is no response to vitamin A administration.
   b. Night blindness—when Bitot's spots cannot be chosen because of their insensitivity as an indicator of vitamin A deficiency.
   c. Keratomalacia—when prevalence of Bitot's spots is very low and yet nutritional blindness is prevalent.

4. Criteria and targets.
   a. A reduction in the prevalence of Bitot's spots by 50% of baseline figures by the end of one year or to below 0.5% by the end of two years, with subsequent maintenance at this level.
   b. A reduction in the prevalence of night blindness by 50% of baseline figures by the end of one year and to below 1.0% by the end of two years with subsequent maintenance at this level.
   c. A reduction in the number of corneal lesions by 50% of baseline, as judged by previously collected hospital data.

   Earlier experiences have shown that in well-implemented projects it has been possible to reach these levels.

   Population target:
   Children between the ages of one and five years:
   a. Universal coverage within a prescribed area when there is evidence that children in practically all regions in the area are at high risk, as defined by World Health Organization criteria.

## Table 2

**Prevalence criteria (in percentage of the preschool-age population, 6 months to 6 years old, at risk) for determining the public health significance of xerophthalmia and vitamin A deficiency**

| |
|---|
| Night blindness (XN) in >1% |
| Bitot's spot (X1B) in >0.5% |
| Corneal xerosis/corneal ulceration/keratomalacia (X2/X3A/X3B) in >0.01% |
| Corneal scar (XS) in >0.05% |
| Plasma vitamin A of <0.35 μmol/l (10 μg/dl) in >5% |

Source: WHO, 1981.

b. High-risk area coverage when through baseline surveys areas which are at high risk can be identified. A suitable administrative unit for coverage has to be defined (e.g., a primary health care unit or subunit).

5. Evaluation design:
   a. Because prevalence rates of indicators may vary widely between areas, a non-intervention versus intervention design cannot be used for evaluation. The design of choice is, therefore, a before and after comparison of the indicators of outcome.
   b. Adequacy of implementation: data to be collected include:
      (1) Sufficiency and regularity of supplies of vitamin A at the distribution points, and facilities for proper storage.
      (2) Extent of coverage—both the proportion of children that received one or more doses and the proportion that regularly received all the doses.
      (3) Records showing details of children, date of administration of the various doses, development of toxicity after dosing, entry of new children into the program at age one, and exit of children at age five.
      (Sources for above three: Health centers in charge of distribution.)
      (4) Understanding by the functionaries at all levels concerning the details of the intervention and its philosophy.
      (5) Community awareness of the intervention with cross-checking of a subsample of children's mothers regarding the actual distribution of supplemental doses.
      (Source: Specific studies by program managers.)
   c. Adequacy of outcome:
      (1) Comparison of baseline prevalence of the indicator with that seen after one/two years of program implementation.
      (2) If baseline data has not been collected, the slope ratio method as described in the IVACG report *Periodic Large Oral Doses of Vitamin A for the Prevention of Vitamin A Deficiency and Xerophthalmia: A Summary of Experiences* can be used, despite some of its limitations.

6. Types and sources of data:
   a. Baseline information on the prevalence of the indicator selected (e.g., Bitot's spots and night blindness from surveys, and keratomalacia from hospital records).
   b. Input data:
      Material:
      (1) Number of doses of vitamin A available for distribution
      (Source: Program managers and health centers.)

Manpower:

(1) Number of workers trained to distribute vitamin A and keep records.

(2) Number of workers trained to supervise distribution and check records.

(3) Number of higher level executives trained to oversee the program and carry out re-orientation courses.

(Sources for manpower data: Program managers and health centers.)

c. Monitoring:

(1) Record all aspects of distribution including incidental occurrence of signs of hypervitaminosis A.

(2) Record all extraneous inputs that could influence baseline prevalence data (e.g., supplementary feeding with or without foods rich in vitamin A, outbreak of an epidemic, improved primary medical care)

(Source: Health centers.)

d. Output:

(1) Number of children who received the vitamin A as prescribed.

(Source: Health centers.)

e. Outcome:

(1) Reduction in the prevalence of Bitot's spots and night blindness; and hospital-based data on keratomalacia by the extent programmed for.

(Source: Repeat survey.)

## Table 3

**Summary of interventions, with suggested targets and criteria[a]**

| Intervention-indicator | Target population | Suggested criteria for prevalence of indicator | Suggested target time period |
|---|---|---|---|
| **Nutrition education program** | | | |
| 1. Duration of breast feeding at least 18 months | Women with children age 0–3 years | Increased to 75% | 5 years |
| 2. Retinol equivalent intake of less than 50% RDA (including breast milk) | Children age 6–36 months | Reduced to 2.5% or reduced by 50% of preintervention level | 5 years<br>1 year |
| 3. Introduction of carotene-rich solid foods by 12 months | Children age 6–36 months | Increased to 75% | 5 years |

*(continued on next page)*

## Table 3 *(continued)*

| Intervention-indicator | Target population | Suggested criteria for prevalence of indicator | Suggested target time period |
|---|---|---|---|
| **Horticulture program** | | | |
| 1. Participation with adequately developed home garden | Families/households with land-use rights | 75% | 2 years |
| 2. Production of specified carotene-rich foods in home gardens for<br>a) at least 6 months per year in semi-arid areas, or<br>b) 12 months per year if water supply is available | Families/households with land-use rights | 75% | 2 years |
| 3. Consumption at least 3× per week of specified carotene-rich foods from home gardens<br>a) at least 6 months per year in semi-arid areas, or<br>b) 12 months per year if water supply is available | Persons in families/ households with land-use rights | 75% | 2 years |
| 4. Retinol equivalent intake of at least 50% RDA | Women of child-bearing age in families/households with land-use rights | 75% | 2 years |
| 5. Retinol equivalent intake of at least 50% RDA (including breast milk) | Children age 6–36 months in families/ households with land-use rights | 75% | 5 years |
| **Food fortification** | | | |
| 1. Serum retinol <10 μg/dl | All ages (0–3 years least likely to benefit) | Reduced to 5% | 2 years |
| 2. Serum retinol <20 μg/dl | All ages (0–3 years least likely to benefit) | Reduced to 15% | 2 years |
| 3. Breast milk retinol <20 μg/dl | Lactating women | Reduced to 15% | 2 years |
| **Large periodic dose administration** | | | |
| 1. Bitot's spot | Children age 12–60 months | Reduced to 0.5% or reduced to 50% of preintervention level | 2 years |
| 2. Corneal lesion | Children age 12–60 months | Reduced to 0.01% or reduced to 50% of preintervention level | 2 years |

[a]For suggested time frames, see text under appropriate intervention.

# V Designs for Evaluation Studies and Sample Surveys

## A. Introduction

Evaluation of vitamin A intervention programs, including general but quite relevant emphasis on epidemiological, statistical, and sample survey matters, has been discussed in earlier IVACG publications. Some of that material is reemphasized here for the convenience of those who may wish to use this manual as their primary guide to evaluation. This chapter expands upon the previous material with additional concepts, detail, and examples so that it may serve as a resource for many aspects of planning and evaluating vitamin A intervention programs. Nevertheless, those responsible for evaluation must take proper account of and adjust for the specifics of the local situation. A statistician or epidemiologist with experience in sample surveys—preferably in the country or locale where the intervention will be done—should be a prominent member of the project team. This will insure appropriate implementation and evaluation and emphasize sound epidemiological, statistical, and sample survey methods.

Evaluation may be required because the intervention is untried, because its efficacy has not been documented, or because the intervention is in a new setting or on a larger scale. In the first two situations, emphasis is more on research and new knowledge, and in the third the interest is in whether the intervention can be adequately implemented or delivered with results as suggested by earlier studies. Although components and methods of quantitative evaluation are similar for all of these situations, the emphasis here is on the latter.

The credibility of the results of an intervention program depends on application of appropriate statistical methods within the context of the overall study plan. While not all of the elements of the study plan are part of the statistical design, the care with which they are done will affect the quality of the results.

## B. Study Designs

The choice of statistical design for evaluation of the effect of an intervention will depend on the type of intervention, the availability and quality

of demographic and pre-intervention status data describing the target population, the feasibility of obtaining additional demographic data, the feasibility of conducting sample surveys, and country or local requirements concerning the scope of the intervention.

The objective is to establish with as much certainty as possible whether the vitamin A intervention is producing its intended effect or outcome. This requires measuring the magnitude of the effect while ruling out other explanations for the results. These other explanations, also called confounding factors or competing explanations, may be very specific to a given intervention program but include some general processes outlined below.

1. Natural or spontaneous recovery from a condition may occur, e.g., Bitot's spots, in addition to recovery influenced by vitamin A intervention.

2. Long-term community trends may produce changes which inflate or inhibit the effect of intervention, e.g., adverse weather conditions that reduce the normal availability of nutritious foods.

3. Short-term events may also inflate or inhibit change, e.g., disruption of normal food delivery systems by strike or natural catastrophe.

4. Different age groups may have differential responses to the intervention.

5. Participation in the intervention may be related to propensity to change, e.g., high-risk children (those who are ill) may not come forward to receive vitamin A capsules.

6. The presence of an intervention program may bring other health-improving benefits, e.g., the increased presence and activity of community health workers giving vitamin A capsules may promote additional access to general health care or to nutrition education.

The influence of confounding effects may be accounted for through the use of a **control group,** that is, potential but untreated targets for the intervention who are otherwise comparable to persons participating in the intervention. Comparability between intervention and control groups means they should be similar in composition (age, sex, etc.), in experience or time-related processes, and in disposition toward participation. (If a control group cannot be constructed or confounding effects cannot be sufficiently removed, statistical adjustments may be used.)

The outcome of an intervention is a change in the level of a measurable variable, such as a change in the percent of children with Bitot's spots. Several common **statistical designs** for measuring outcome of vitamin A interventions and their relationships may be described with the aid of Figure 1, with the indicator prevalence of Bitot's spots as an example. $I_{BA}$ is the change in prevalence in the intervention group, from before to after intervention, without benefit of information on confounding factors. $A_{CI}$ is the difference in prevalence between the intervention group and the control group. $I_{BA} - C_{BA}$ is the difference between the intervention group's change in prevalence and the control group's change in prevalence during the same time period.

In a "true" experimental design (a **randomized experiment** or trial), intervention is randomly given to or withheld from persons or units of the target population to create the intervention and control groups; $I_{BA} - C_{BA}$ measures the net effect of intervention; and effects of confounding factors will be minimized. Randomized trials are usually restricted to small-scale projects covering only a part of a potential target population; are usually not feasible in areas where an intervention has been in place for some time; may be unacceptable on political, ethical, or legal grounds; and are expensive.

Approximations or alternatives to the randomized design, or "quasi-experiments," can provide results free from many confounding processes if they are properly conducted. First, **constructed controls** may be used in situations where the intervention program has already begun or where intervention must be delivered to all members of a designated unit (e.g., all children in a village). These controls are selected so that they are comparable in essential respects to the intervention group, e.g., intervention is withheld from children in a comparable village. Practical considerations usually lead to matching groups on aggregate characteristics rather that on individual characteristics, e.g., matching on average age rather than requiring the same number of three-year-old children in both groups.

Characteristics useful in devising constructed control groups include those descriptive of individuals (e.g., age, sex, education, and health parameters such as height and weight), of households (e.g., number of children, socioeconomic status, ethnicity), and of communities (e.g., population size, land area, urban/rural). $A_{CI}$ measures the effect of intervention for these concurrent controls. A variation in this design, but one with more difficulty in interpretation of results, includes **historical controls** whose characteristics precede the period of the intervention program and may be known.

Second, persons may serve as their own controls in a **before-and-after** intervention design, where $I_{BA}$ measures the effect of intervention. Time-related confounding factors are especially relevant here, including the simple fact of change in age during the program.

In non-experimental designs, statistical adjustments or **statisti-**

**cal control** may be used in the common situation where control groups cannot be constructed or when the design is insufficient to remove confounding effects. This may be done through a cross-sectional study or survey, in which those who did and did not participate in (or receive) the intervention are identified, and information on possibly confounding variables is collected. Adjustment for relevant differences between participants and non-participants is done by statistical techniques. The effectiveness of this adjustment is limited by the need for prior knowledge of relevant confounding variables. However, the effects of self-selection (differential reasons for participation and non-participation) are usually not amenable to removal from estimates of intervention effect.

Actual implementations of intervention programs permit and benefit from a combination of control groups and statistical control as a mixed strategy, but this approach requires special skills in statistical analysis and access to computers.

A non-experimental design of a distinctly different nature is the **case-control design.** While the designs mentioned so far proceed from cause to effect, the case-control design proceeds from effect to cause. Cases of the disease or condition (e.g., children with xerophthalmia) are compared with non-cases or controls (e.g., children without xerophthalmia) with respect to a current or previous study factor level (e.g., receipt of vitamin A supplement). To obtain an odds ratio, the odds of cases having received vitamin A is divided by the odds of controls, who have also received vitamin A. This **odds ratio** approximates the relative rate of the disease (xerophthalmia) among those who were exposed to the factor (e.g., vitamin A supplement) compared to the non-exposed (e.g., did not receive vitamin A supplement).

A major advantage of the case-control design is the smaller sample sizes required when prevalence of disease is low. Selection of representative cases and controls—on the basis of their disease status, rather than their intervention status—must not be influenced by their intervention status. A disadvantage of this method is that the result is reported as a ratio of odds, or approximately as a ratio of risks or proportions, rather than as a difference between proportions which is more familiar and more readily interpreted. The case-control design is displayed in Figure 2. Statistical control or adjustment for confounding factors is also possible here.

## C. Sample Survey

Evaluation of vitamin A intervention on a moderate to large scale, as in regional or national programs, requires one sample survey for the constructed or concurrent controls design and for the cross-sectional statistical control design, and two sample surveys for the before-and-after design and for the randomized trial.

The principal steps in a sample survey are as follows:

1. **Objectives of the survey** should be clearly stated and should guide the details of planning the survey.
2. The **population to be sampled** is defined and generally should coincide with the population about which information is wanted (the target population).
3. **Data to be collected** should be relevant to the purposes of the survey but should omit no essential data.
4. The **degree of precision** desired must be decided.
5. The **method of measurement** is chosen (e.g., interview or examination) and the survey instrument or data collection form must be prepared.
6. The **sampling frame,** a list of units to be sampled, is described or created.
7. The **sampling plan** is chosen, with sample size estimated from the properties of the plan and from knowledge of the degree of precision required.
8. **Pretest** the survey instrument and field methods.
9. **Field work** is organized and then carried out.
10. **Summary and data analysis** are done, including editing of data forms.

A **sampling plan** is a rule for drawing a sample from a target population, with a procedure for making estimates of population characteristics from the results of the sample. To be useful, an estimate must be representative of the entire target population and must be sufficiently close to the true but unknown population characteristic.

**Stratified, multistage, cluster sampling** is recommended as a practical and effective sampling scheme in evaluation of vitamin A interventions. To begin, the target population is divided or stratified into a number of parts called **strata** on the basis of some convenient characteristic so that members of the stratum are homogeneous for that characteristic (e.g., urban and rural, large and small provinces). Stratification usually improves and rarely reduces sampling accuracy. In the **first stage,** each stratum is divided into **primary units,** which are convenient administrative units of moderate size (e.g., districts), and a sample of these primary units is taken. At the **second stage,** a sample of **secondary units** (e.g., villages) is taken within each selected primary unit. If villages are small, the process may stop here, with villages being the final cluster-unit selected. Often there will be a **third stage,** in which a sample of households is taken within each village selected. Probabilities of select-

ing primary, secondary, and tertiary stage units can be determined, which may simplify calculation of results. These and other concepts of sample design are expanded in the Appendix.

Although establishing a sampling plan and estimating sample size requirements are very important, the other steps also require close attention, as the quality of the survey and ultimately the credible measurement of the effect of intervention depend on them. The survey instruments must be carefully developed, tested, prepared in the local languages, and printed in a format which can be effectively used by the field team and in computer entry and data analysis. While examinations and interviews should be as brief as possible, they should also record sufficient information on confounding variables to permit statistical adjustment for group comparability. Field teams must be adequate in size, number, composition, and training. Logistic support must be established, e.g., transport and daily support of the field team, and printing, distribution, and collection of data forms. Residents of a survey area should be informed of the purpose of the survey, and the cooperation of village leaders and local health workers must be obtained. Field teams must follow closely the sample selected and the prescribed procedures, to avoid introducing additional confounding and inaccuracies.

## Figure 1

**Indicator (prevalence of Bitot's spots)**

| | Before Intervention Program | After Intervention Program | Difference |
|---|---|---|---|
| Intervention Group | $I_B$ | $I_A$ | $I_{BA} = I_B - I_A$* |
| Control Group | $C_B$ | $C_A$ | $C_{BA} = C_B - C_A$ |

$B_{CI} = C_B - I_B$ $A_{CI} = C_A - I_A$†

Program Net Effect $= I_{BA} - C_{BA}$‡

$I_B$, $C_B$ = prevalence of Bitot's spots before intervention program is begun, for intervention and control groups, respectively.

$I_A$, $C_A$ = prevalence of Bitot's spots after intervention program is completed, for intervention and control groups, respectively.

*Used in nonrandomized "before-and-after" designs.

†Used in nonrandomized designs with constructed controls.

‡Used to measure intervention effect in randomized trials.

## Figure 2

Case-control design: Sample of $n_1$ cases and $n_2$ controls

| | | Cases* | Controls |
|---|---|---|---|
| Exposure† | | | |
| | Yes | a | b |
| | No | c | d |
| Total | | $n_1$ | $n_2$ |

OR = odds ratio = (a/c)/(b/d) = ad/bc‡

*Cases are persons with the disease or condition, e.g., Bitot's spots; controls are persons without the condition.

†Exposure here means receipt of vitamin A, or participation in vitamin A intervention.

‡OR < 1 means intervention is effective, i.e., reduces risk of Bitot's spots.

# Appendix[1]

## Introduction

The purpose of this section is to familiarize those wishing to conduct a household survey to evaluate a vitamin A intervention program with the main issues in sample design. Familiarity will allow survey directors and managers to collaborate more effectively with sampling experts in the design and selection of the sample, and also allow them to conduct some sample selection for small surveys. Large scale surveys will require the expertise of a sampling statistician in most, if not all, phases of the design and selection.

## 1. Sample Design

Sampling design has perhaps the most highly developed theory among the various aspects of the survey design process. Nonetheless, sample design, as with many design activities, calls for the exercising of a considerable degree of judgment in deciding the appropriate choice of design features best suited for a particular problem. The need for expert judgment makes the development of a recipe for generic sample designs difficult to specify, even for a specific problem such as the evaluation of vitamin A interventions in a population. Sample design is highly dependent on the setting of the survey and the types of information available for the population to be surveyed. There is no sample design cookbook that adequately anticipates the diverse survey settings and populations for which a sample is needed.

The section begins with a description of the major steps in the sample design and selection process for a household survey. This is followed by a discussion of several sample design principles and the description of four general features of every sample design: population specification, consideration of design constraints, sampling frames, and stratification. An example of a multistage sample design selected with

[1]This Appendix was prepared by James M. Lepkowski, Ph.D., and Roy C. Milton, Ph.D.

varying probabilities of selection then follows. This example is intended to illustrate the basic techniques that are needed for sample selection in a developing country. Finally, there is a description of how estimates of proportions or rates and their standard errors can be calculated from data collected in the example design.

**Overview of the Sample Design Process.** The basic problem of sample design is to find a design which minimizes the error of the estimates from the survey for a given fixed cost of conducting the survey. The design process in some ways resembles the economic problem of resource allocation: how to allocate the limited resources of the survey (*i.e.*, costs) in order to minimize the error of the survey results. The cost of controlling one source of error in the survey results must be compared and possibly traded against the cost of controlling another source of error.

The design must take account of additional constraints on the survey besides cost: geographic, organizational, and personnel constraints which are part of the overall survey problem. Sampling theory can be used to guide the choice of design options which will minimize the error under these various constraints, but lack of quantifiable information often requires the exercise of judgment by a sampling expert to decide which combination of design features is likely to lead to the most efficient sample design.

There are several steps in the sample design process, many of which overlap with those in the general survey design process. These steps are not necessarily sequential in nature, but most survey designs will follow each step in approximately the order indicated here.

The first step is to **specify the survey objectives** clearly and fully. This step likely will include specification of some limited number of variables, or even a single variable, for which the error of the survey is to be minimized. Since the survey is likely to include a large number of variables collected simultaneously, minimization of error for all variables collected in the survey will not likely be possible. In addition, knowledge about the error and cost parameters needed for the minimization algorithms derived in sampling theory is likely to be available only when a similar survey has been done before in the same location or country. Often the detailed knowledge needed to perform the minimization will simply not be available. As a result, practical sample design is sometimes completed with limited use of the theoretical minimization techniques available. Fortunately, modest departures from optimum solutions often do not result in major losses in design efficiency, and nearly optimal solutions can be used without large losses. Some rough, "back-of-the-envelope" calculations based on theoretical methods may be employed to obtain a design that is at least near optimum for a given problem.

Closely related to specification of survey objectives are two

other steps: the **specification of the population** and a **consideration of design constraints**. An intent to examine children in a program intervention area must be translated into detailed specifications about the nature and extent of the population. The population specification, as well as other design features, may also be modified given the constraints under which the sample must be designed.

Often at the same time as these specification and constraint consideration steps, the materials for the sample selection are being collected. Referred to as the **sampling frame**, these materials may include maps, lists of units of various types, and forms for collection of data about other types of units. Once some or all of the sampling frame is collected, **specification of sample selection techniques** can take place. The techniques typically include stratification, varying probabilities of selection, and multistage sampling. At some point during the survey, there will be a need to make a **selection of households and persons**, probably performed by survey staff in the field. Finally, the sample design process includes **data processing and estimation** activities at the conclusion of data collection.

Thus, the sample design process is an integral feature of the entire survey from specification of objectives and populations to final estimation and report writing. Before describing several of these steps more fully, it is useful to discuss several sample design principles that underlie much practical sample survey design.

**Sample Design Principles.** Samples selected haphazardly or samples of convenience are not preferred by survey practitioners because it is difficult to control many types of error in such surveys. It is also impossible to estimate sampling errors for them without strong assumptions about the nature of the population. A preferred approach is **probability sampling**, the selection of a sample in such a way that every population element has a known nonzero chance of being selected. When probability sampling methods are used, the error of sample estimates that occurs because only a sample and not the entire population has been examined can be estimated directly from the sample results alone. Probability sampling requires careful application of the principles of sampling theory to assure known and nonzero chances of selection for each population element. It is one of several reasons why the expertise of a sampling statistician should be sought in the design of sample surveys.

Given that each population element will have a known nonzero probability of selection in such designs, the practitioner can introduce unequal probabilities during the course of selection in order to "over–" or "under–sample" designated population groups. This over– or under–sampling may be useful for providing larger subsamples within the overall sample of important population subgroups for more detailed study. Departure from **e**qual **p**robability of **se**lection **m**ethods (often referred to as **epsem** designs) requires that adjustments be

made in analysis when subgroups which have been sampled at different probabilities are combined to form overall estimates. The adjustment procedure is made through the use of weights applied to survey observations. Weighted estimates are more difficult to make than unweighted estimates, and survey practitioners generally prefer to use epsem designs and avoid weights in analysis.

Sampling for populations with an available list of each element of the population can be a straightforward task. However, household surveys in most countries of the world do not have the luxury of a list of households from which to select the sample. Worse, it would be prohibitively expensive to create such a list just for the purposes of sample selection. One procedure which can be used in household surveys when lists of households are not available is **area sampling**. Instead of sampling households directly, a list of areal units which covers the geographic extent of the target population completely is assembled as a sampling frame. The areal units are themselves selected, and households within selected areal units are then listed and selected for the survey. Lists and maps of such areal units are often readily available from a recent census, and there are typically counts of households or similar units for each areal unit.

Area sampling is usually combined with **multistage selection** to introduce further economies in sample selection. Once a selection of areal units has been made, further sampling activities are conducted only within selected units. The further sampling may involve selection of additional areal units within the selected areal units, or it may consist of creating lists of households within the selected areal units, lists from which a final sample of households can be selected. Each successive stage of sampling after the first is limited to units selected at a previous stage, reducing the amount of listing of units that much be done at each stage.

**Population Specification.** The specification of the population to be surveyed is essential to the proper conduct of a survey. It is an activity which requires a good deal more effort than may be at first apparent. The population units and extent of those units must be specified together with the content of the survey measurements which might be applied to them. For example, a survey to evaluate a vitamin A intervention program may specify a population consisting of all children (the units) ages 6 months to 10 years residing in a four province program target area (the extent of the population). A further specification would be that the children's physical condition will be examined through such devices as height and weight measures, plus blood will be drawn for serum vitamin A level assessments.

The survey may have already specified in the survey objectives that estimates be produced for key subgroups. The population specification should clearly indicate the need for additional consideration of certain subgroups of the population.

The population specification must further specify the time period during which the observations are to be made. This may also indicate a date for which inferences about the population are to be made. For example, the children may be examined over a specified four month period required for completion of data collection, while findings are intended to refer to the entire population of children as of the midpoint of the data collection period, despite the four month data collection period.

When the actual data collection takes place, survey teams may find that some of the eligible population who usually reside in sample households are not present at the time of the survey. Further, at the time of the survey they may find guests who are not usual residents of the household. Two definitions are used to assist survey teams in determining who is and is not to be examined: *de facto* and *de jure*. The *de facto* population is the one currently living in the household, whether it is their usual residence or not, while the *de jure* population is the usual residents of the sample households. For most surveys, the *de jure* definition is preferred, because it avoids potential double counting of persons that may occur under the *de facto* definition. Under the *de jure* definition, persons who are not present at the time of examination are considered nonrespondents, while guests are considered to be ineligible for the survey.

**Design Constraints.** Constraints on the survey may be imposed from a variety of sources. These include a limited budget, a limited time period for data collection, and the quality of the personnel to be used for data collection and processing. Each will have an impact on many different aspects of the survey.

For example, the sample size will be greatly influenced by the budget available for the survey. The larger the total sample, the more precisely will survey estimates reflect actual population values. But the survey may also need to provide precise estimates for subgroups of the population. There are two basic approaches to handle greater precision for population subgroups: larger overall sample sizes or an over–sampling of the subgroup itself. The extent of the over–sampling will be a major determinant of which of these two approaches will be applied in a given situation.

Sample size will also be influenced by the joint effect of a fixed budget, a limited data collection period, and the quality of the personnel. For example, to the extent that the personnel are not well–suited to careful data collection under field conditions, errors may occur which overwhelm the sampling error for all but the smallest samples. A vitamin A intervention program worker employed to list households for an evaluation survey may not be aware of the importance of listing all households in a given areal unit. Through oversight, or even deliberately avoiding hard to reach portions of an areal unit, the sample selection can be seriously biased by a poorly motivated or

trained staff member performing low quality work. A smaller sample is preferred in such situations, with the budget saved by collecting data from fewer subjects used to improve the training and supervision of the survey staff.

In many situations, seasonal events may constrain the period available for field work. Data collection must often be completed between two rainy seasons, or important holidays such as harvest festivals may make data collection virtually impossible during portions of the year. At the same time, a survey organization may want to spread the data collection over an extended period to limit the number of field staff needing training or being borrowed from other project activities at any given time. Care must be taken to assure that scheduling does not cause the effects of seasonal trends in disease occurrence or seasonal migration to bias survey results.

**Sampling Frames.** The choice of specific sample selection techniques is perhaps most importantly influenced by the types of materials available for sample selection. The materials may consist of lists of units such as large and small geographic areas, addresses, households, persons, or registered or eligible voters. There may be in some settings a large variety of materials from which to choose, while in others the choice may be extremely limited if not nonexistent. There are a number of characteristics which should be considered in choosing or developing a sampling frame for a survey.

The materials should be complete, covering all units in the population of interest, and they should be as up to date as possible. Each unit on a list should only be listed once, since duplicate entries allow a given unit multiple chances of entering into the survey. Further, removal of duplicates can be an expensive and difficult task. The units should be clearly delimited to avoid confusion in the field about whether subunits are to be included in one unit or another. For example, to reduce confusion about whether or not a particular housing unit is contained within an areal unit, the boundaries of the unit should follow relatively permanent physical features that are readily recognizable in the field. Finally, materials which provide ancillary data about the units which might be useful in the sample design are preferred. Maps of areal units are much more useful if there is accompanying population and other data about each unit in the population.

For area sampling, the sampling frame will typically consist of maps and accompanying data about the geographic units defined by the maps. There may be several levels of units defined by these maps: provinces, districts, subdistricts, villages. A choice must then be made about which of these units are to be used in sample selection and which may be ignored or skipped in the selection process. This choice often requires careful judgment about the relationship between the size and other characteristics of the units and the sampling and other errors which will arise in the final estimates.

Ultimately, these materials will provide a set of areal units from which the sample of households and eligible persons must be selected. The selection of households and persons will require the creation of lists in the field. The amount of listing that is to be done depends on design constraints such as the quality of survey personnel, the survey budget, and whether the same materials might be used for more than one survey. Lists of housing units may remain useful over extended periods of time, but lists of persons will become out of date rather quickly. In most cases, the listings will be returned to a central location where checks can be implemented and a final selection made by trained personnel. Lists of selected households, together with clear instructions on how to handle problem cases, can then be returned to the field for survey data collection.

Not all data collection activities may need to return listings to a central location for processing. It would be poor survey practice, though, if a list of some type were not prepared at the last stage of selection. For purposes of quality control alone, careful listing of households or persons must be completed at some stage of the sample selection process.

Available materials from a recent census may not provide units that are small enough for listing purposes. In this instance, area sampling may involve the creation of areal units, a process that requires consideration of such issues as clear demarcation of unit boundaries, geographic size of the unit, and the availability of ancillary data for the units. These considerations involve much of the "art" in sample design and survey practice.

**Stratification.** Stratification is a sample selection technique in which a population of units is divided into subgroups or strata, and a sample is selected independently from each stratum. The strata must be mutually exclusive and exhaustive, assigning every unit to one and only one stratum. Since stratification is typically a straightforward and simple process to use, sample designs universally attempt to employ stratification at some point in the sample selection process.

Stratification has several purposes. For one, to the extent that units within strata are similar with respect to the characteristics under study, stratification can lead to reductions in sampling error. It is the within stratum variability which is used to estimate the sampling error of estimates from the sample; within stratum homogeneity reduces the size of the sampling error of estimates. Since the characteristic of interest will not be available for creating strata, characteristics that are closely related to those of interest are used to form the strata instead. Another purpose of stratification is to assure that different subgroups of the population will appear in the sample, and that they are properly represented in the sample. Random selection without stratification may include by chance alone too few or too many of a given subgroup. With stratification, the proportion of the subgroup in the sample can

be controlled to be the same as in the population, or the sample proportion may differ from the population proportion to reflect study specific design constraints.

For a sample of children ages 6 months to ten years, stratification by gender may be employed at the last stage of selection to assure that approximately equal numbers of males and females are included in the sample. But stratification can be applied at many other points in the sample selection as well. For example, in area sampling the areal units may be divided into subgroups, and separate samples of those units are chosen from each stratum. The same basic principles apply: mutually exclusive and exhaustive subgroups of similar units are created, and from these independent selections of units are made.

The direct or explicit creation of subgroups is not the only method for stratification. Stratification may be employed implicitly in sample selection as well. For example, a list of units may be ordered with respect to some characteristic related to the study measures of interest. That is, units that are similar to one another with respect to that sorting characteristic (and presumably the related study measure as well) are placed next to one another on the list. A systematic sample is then selected from the ordered list in which every *k*th unit is selected after a random start. The combination of an ordered list and systematic selection is an implicit stratification of the population during selection.

**Varying Probability and Multistage Sampling.** In area sampling, the areal units are typically of varying population size. For instance, one unit may be five times larger than another; proportionately selected samples from the two units will also have the five–fold variation in size. A technique for handling the variation in size of units is to employ the size in the sample selection process: **P**robability **P**roportional to **S**ize or **PPS** sampling. Thus, the larger unit is assigned a probability of selection that is five times larger than that assigned to the smaller unit. Of course, population elements in the larger unit then have five times the chance of being selected as those in the smaller unit. Some adjustment must be made to account for this discrepancy.

One way to correct for these unequal probabilities of selection is to sample within the units at rates that are inversely proportional to those used to select the unit. Thus, the larger unit would be subsampled at a rate that is only one–fifth as large as that used to subsample the smaller unit. If the next stage of selection were households, then the households of both the larger and smaller unit would be sampled at different rates to restore epsem sampling. In addition, this two stage varying probability of selection design also provides a sample with the same number of households selected from each unit. The organization of field work will be easier when the workloads for units of very different sizes are identical.

Selection with PPS requires careful control of the selection probabilities at each stage of sample selection. A common error in the

selection is to assume that the final stage of selection requires that a fixed number of units be selected. The PPS selection requires that the final stage be selected at a rate which, if the size of the units is known exactly for every unit, will lead to equal numbers of subunits being subsampled from each unit. Unfortunately, PPS selection must often be conducted with measures of size that are somewhat out of date or otherwise inaccurate. For instance, size measures may indicate that one unit is five times larger than another when in fact it is actually ten times larger. Using PPS sampling methods, the subsample from that unit will not be identical in size to those selected from every other unit; it will be twice as large. Thus, PPS selection will not insure that the workloads are exactly the same across units, but, as long as the size measures are reasonably accurate, the subsample workloads will be similar across units of varying size.

The principles of PPS selection are commonly applied in multistage sample designs. In order to describe the PPS selection procedure in a clear and concise manner, we must introduce statistical notation. However, before doing so it will be useful to describe a hypothetical sample design for a vitamin A intervention evaluation program that can later be used to illustrate the meaning of the statistical symbols.

Thus, to fix these ideas more firmly, suppose that we want to select a sample of children ages 6 months to ten years from a single district of a developing country for purposes of studying the effectiveness of a vitamin A intervention program being conducted in the district. In particular, suppose that the district is composed of the 24 subdistricts shown in Table 1, and that we have the hypothetical count of households in each as of the last census. (The data in Table 1 are actually drawn from census and other materials from the Kingdom of Nepal; size measures have been altered to fit the purposes of this illustration.) A list of all of the villages and their household counts in the 24 subdistricts is not available from one source, although lists of villages and their household counts are available in the administrative office for each subdistrict. Thus, it is proposed that a sample of subdistricts be selected, and then villages subsampled within selected subdistricts. Finally, households will be listed in each selected village, a sample of households selected, and all eligible children within selected households will be chosen for examination. This represents a three-stage sample of households: subdistricts, villages, and households. One may consider the "selection" of all eligible children within selected households to be a final, fourth stage of selection in which eligible children are selected with certainty. For our purposes, we do not consider this last step in the selection process to be a stage of sampling.

Suppose that a sample of 450 examined children is needed to provide enough observations to obtain reasonably precise estimates of important characteristics for the evaluation. It is expected that 90 percent of the selected children will cooperate with the study staff and

## Appendix Table 1

**Hypothetical measures of size for 24 subdistricts in one district undergoing a vitamin A intervention program.**

| Subdistrict no. ($i$) | Subdistrict name | Census household count ($Mos_i$) | Cumulative count | Number of villages |
|---|---|---|---|---|
| 1 | Gantha | 369 | 369 | 7 |
| 2 | Balha | 175 | 544 | 3 |
| 3 | Shreekot | 341 | 885 | 7 |
| 4 | Lui | 237 | 1,122 | 5 |
| 5 | Dhankot | 214 | 1,336 | 4 |
| 6 | Maila | 403 | 1,739 | 8 |
| 7 | Louthi | 292 | 2,031 | 6 |
| 8 | Shree Napre | 301 | 2,332 | 6 |
| 9 | Vihi | 130 | 2,462 | 3 |
| 10 | Jair | 158 | 2,620 | 4 |
| 11 | Nathrapu | 277 | 2,897 | 5 |
| 12 | Kajuri Mahuwa | 221 | 3,118 | 4 |
| 13 | Karle | 237 | 3,355 | 5 |
| 14 | Sheri | 231 | 3,586 | 5 |
| 15 | Rara | 175 | 3,761 | 3 |
| 16 | Pina | 340 | 4,101 | 7 |
| 17 | Karkinada | 290 | 4,391 | 6 |
| 18 | Shree Napar | 247 | 4,638 | 5 |
| 19 | Ruga | 327 | 4,965 | 7 |
| 20 | Rova | 362 | 5,327 | 7 |
| 21 | Nagri | 209 | 5,536 | 4 |
| 22 | Pulu | 102 | 5,638 | 3 |
| 23 | Dolfu | 174 | 5,812 | 4 |
| 24 | Mugu | 163 | 5,975 | 4 |
| | Total | 5,975 | — | 122 |

will be examined; a sample of 500 children is needed to obtain the required number of examinations.

The sample of 500 children could be selected in several differ ways from the 122 villages. One approach would be to visit each household in the study area, list every child, and then select a stratified sample of children from the list. Obviously, this would require a substantial expense for the listing. Further, the sample would be spread throughout the study area, probably with at least one child in each of

the 122 villages. It would be costly to examine a sample of children scattered over all villages in the study area.

Alternatively, the sample could be clustered within villages to reduce the amount of listing and travel required to obtain the examinations. At one extreme, only a few villages might be selected and all age–eligible children in those villages examined for the study. From a cost viewpoint, this design is attractive since listing and travel costs would be limited to a few villages. However, if the precision of the estimates obtained from the study is considered, such a highly clustered design is not desirable. Even for characteristics concerning vitamin A, there is likely to be some positive correlation among children within the same village. For example, in a village in which one child is found to have serum vitamin A levels lower than 15 $\mu$g/dl, there are more likely to be other children with lower values in the same village compared to another village in which such a child has not been found. Given this positive correlation, the more children selected from a sample village, the less new information each child provides us about the study characteristics. The less information each new observation provides, the less precise will be the estimates obtained from the sample. Statistically, this loss of information and precision is assessed in terms of an "intra–class" correlation. The size of this correlation determines the size of the loss of precision of survey estimates.

The distribution of the sample must thus be somewhere between completely scattered across all villages and completely clustered in a few villages. The optimal distribution of the sample across villages depends, on the one hand, on the costs of conducting the survey (particularly obtaining listings of children within and travel to villages) and, on the other, on the size of the "intra–village" correlation among children of important study characteristics. Although beyond the scope of this present discussion, statisticians have devised methods for determining the optimal distribution. For our purposes, let us consider instead a practical approach to determining an allocation of sample to sites. (One might later compare the results of this practical approach to the optimal distribution to determine if it is reasonably close to the optimum.)

Suppose that the examination procedure, which includes a blood draw and extensive physical measurements, limits the number of examinations which can be completed in one day to 25. More than one day in a location is needed to allow at least one night in a location to contact children who failed to be examined the first day. If two examination days are allowed per location, approximately 50 children can be examined per location.

In a single village, it will be easier to list children by household prior to final selection. Thus, we need to determine the number of households to select which will yield the required number of children.

Suppose that in the last census approximately one-third of the population was ages 6 months to ten years and on average there are 6 persons per household. If we examine all eligible children in a selected household, then 50/2 = 25 households containing 6 × 25 = 150 persons are needed for the study in a single location. In the study area, there are 5,975/122 = 49 households per village. On average, approximately one-half of the households in a village are needed to provide enough children for an examination location. In total, 500/50 = 10 villages are needed to obtain the required number of completed examinations.

There is one other feature of the sample selection which needs to be specified: the number of subdistricts and villages per subdistricts to be selected. Cost and precision considerations similar to those for the number of children per village must be examined to choose an appropriate number of villages per district. For example, one could select ten subdistricts and one village per selected subdistrict to spread the sample out across the study area as much as possible to improve the precision of the survey estimates. Alternatively, five subdistricts could be selected and two villages subsampled from each to reduce travel costs at the expense of somewhat less precise survey estimates. For the purposes of the present design, suppose that the decision has been made to follow this latter alternative.

Given these design characteristics, let us turn to a consideration of how PPS selection methods might be used to select a sample of five subdistricts, two villages per subdistrict, and a cluster of approximately 50 children per selected village for this study. We now need to introduce some statistical notation to describe the PPS methods more clearly. Suppose that we have a population of $A$ units and for each one we have a **M**easure **o**f **s**ize, labeled $\boldsymbol{Mos_i}$ for the $i$th unit, which may be a count such as the number of households at the last census. If a PPS selection of $a$ of these units is made at the first stage of selection, the probability of selection of each unit, say $p_i$, can be computed as

$$p_i = \frac{a\,Mos_i}{\sum Mos_i}$$

where $\sum Mos_i$ is the sum of all measures of size for all the units.

Following the illustration, there are $A$ = 24 units in the population from which $a$ = 5 are to be selected using the census household count $Mos_i$ for each district as the measure of size. This measure of size is clearly incorrect at present for a given subdistrict, but the relationship between the census counts among the 24 subdistricts is probably similar to that which might be observed if a complete census were to be conducted at the time of sample selection. Even though the measures are known to be out of date and inaccurate, they can be used

effectively in a PPS selection of subdistricts for the survey. Finally, the sum of the measures is $\sum Mos_i = 5{,}975$.

A frequently employed method for selecting a PPS sample of units is to cumulate the sizes of the units as shown in Table 1, and then select the sample using a systematic sample with fixed interval applied to the cumulated sizes. In particular, if $a = 5$ districts are to be selected with probabilities proportional to their measures $Mos_i$, the interval is calculated as $\sum Mos_i / a = 5{,}975 / 5 = 1{,}195$. That is, one subdistrict will be selected from every cumulation of 1,195 census household counts. To determine which subdistricts are selected, a random number from 1 to 1,195 is obtained from a table of random numbers. In this case, suppose the random number is 519. The cumulative column is examined to find the first subdistrict which has a cumulative total that is greater than this random start, in this case, Balha subdistrict with cumulative total 544. Thus, Balha is the first selected subdistrict, and its probability of selection in this first stage is $p_2 = a\ Mos_2 / \sum Mos_i = 5 \times 175 / 5{,}975 = 0.146042$.

To determine the next selection, the interval 1,195 is added to the random start to obtain the next selection number, 1,195 + 519 = 1,714. The cumulative total column is again examined to find the first subdistrict with a cumulative total that is greater than the selection number, Maila subdistrict. Again, the probability of selecting Maila subdistrict is computed as $5 \times 403 / 5{,}975 = 0.33724$. The selection of subdistricts continues in this manner until the fixed number of five selections has been chosen.

Although it could not occur in this instance, it is possible for a subdistrict to be selected more that once for a sample using this procedure. In particular, if one of the 24 subdistricts had a census population count of more than 1,195 (while the sum of measures of size for the study area remained the same), that subdistrict would be selected at least once and possibly two or more times, depending on its actual size. If it were selected twice, two subsamples would have to be selected from it at the subsequent stages of sampling. Alternatively, all such large units can be identified prior to selection and placed in a separate stratum. Since they will fall into the sample anyway, they can be subsampled directly at some fixed overall probability.

Now let us consider the second stage of selection, villages within selected subdistricts. It is important to distinguish the *conditional probability* of selecting a village, given that the subdistrict has already been selected, from the *overall probability* of selecting the village. Let $p_{1i}$ denote the probability of selecting the $i$th subdistrict. (This was previously denoted as $p_i$, but the subscript 1 has been added to indicate selection of subdistricts at the first stage.) Now let $p_{2ij}$ denote the probability of selecting the $j$th village within the $i$th subdistrict, once the $i$th subdistrict has been selected. The overall probability of selecting the $j$th village in the $i$th subdistrict is then:

$$p_{ij} = p_{1i} p_{2ij}.$$

Similarly for three stages of selection, the overall probability of selecting the third stage *ijk* units (households in this example) is denoted as

$$p_{ijk} = p_{1i} p_{2ij} p_{3ijk},$$

where $p_{3ijk}$ denotes the conditional probability of selecting the *ijk*th third stage unit given that the *i*th first stage and *ij*th second stage units had already been selected.

In PPS selection, the key step is determining these conditional probabilities in such a way that the overall probability of selection of the final stage units $p_{ijk}$ is equal for all third stage units. This overall constant sampling rate for final stage units we shall denote as $f$. In the example, we seek a sample of 500 eligible children in a population of 6 × 5,975 = 35,850 persons, one–third of whom are expected to be eligible, or 35,850 / 3 = 11,950 children ages 6 months to ten years. Hence, $f$ = 500 / 11,950 = 1 / 23.9 is the desired overall sampling rate. (Alternatively, the sampling rate could have been computed in terms of households as 500 / 2 = 250 households to be selected from a population of 5,975.)

Of course, this sampling rate will be appropriate if the total population is currently still equal to 35,850 in the study area. More likely there has been some population growth since the last census, and if we apply this overall sampling rate to the increased population, a sample larger than needed will be obtained. For example, suppose that it is expected that the population has increased in the country as a whole by two percent each year since the last census eight years ago. Over the eight year period, the population in the study area would be expected to increase by a factor of $(1.02)^8$ = 1.17 or 17 percent to a total of 1.17 × 35,850 = 41,945. That is, as of the time of the survey we expect a total population in the study area of approximately 42,000, of whom 42,000 / 3 = 14,000 will be eligible children. An overall sampling rate of $f$ = 500 / 14,000 = 1 / 28 is needed to obtain the required sample of 500 eligible children. Given the uncertainty about population growth rates and the proportion of the population which is eligible for the study, one might prefer to use the previous overall sampling rate of 1 / 23.9 to obtain a larger sample than needed. In this example, though, let us fix the overall sampling rate at $f = p_{ijk}$ = 1 / 28 for the final stage of selection.

In the second stage of selection, which occurs within selected subdistricts, the conditional probability of selecting the *ij*th village is similar to the first stage probability:

## Appendix Table 2

**Hypothetical measures of size for eight villages in the selected subdistrict Maila.**

| Village no. ($j$) | Village name | Census household count ($Mos_{ij}$) | Cumulative count |
|---|---|---|---|
| 1 | Thadijhinja | 74 | 74 |
| 2 | Duhaubi | 53 | 127 |
| 3 | Aaurahi | 55 | 182 |
| 4 | Ekrahi | 61 | 243 |
| 5 | Dhabouli | 58 | 301 |
| 6 | Machijhitkaiya | 21 | 322 |
| 7 | Fulbariya | 50 | 372 |
| 8 | Kamala | 31 | 403 |
| | Total | 403 | — |

$$p_{2ij} = \frac{a_i Mos_{ij}}{\sum_j Mos_{ij}}$$

where $a_i$ is the number of second stage units to be selected within selected first stage units, $Mos_{ij}$ is the population count or size of the $ij$th second stage unit, and $\sum_j Mos_{ij}$ is the sum of size measures for the villages in the $i$th first stage unit. In the example, the same number, $a_i = 2$, of villages are to be selected within each selected subdistrict.

Table 2 presents the villages and their household counts for one of the selected subdistricts, Maila. Since $a_i$ has been chosen, and the $Mos_{ij}$ are available, the selection of second stage sampling units can be made in a manner identical to that used to select first stage units. Thus, for two village selections from the $\sum_j Mos_{ij} = 403$ total measures across the eight villages, an interval of $403 / 2 = 201.5$ can be used to select villages with PPS.

A small complication arises in this instance since the interval is not an integer. Systematic selection with fractional interval is possible, although other approaches can be used in this situation as well. A random start is chosen from 1 to 2015, and then a decimal point is

inserted before the last digit. Suppose the random start was 348. Inserting a decimal point before the last digit provides the random start of 34.8. The selection number to be applied to the cumulative column is obtained by dropping the the decimal portion of the selection number, 34 in this case. The first village, Thadijhinja, with cumulative total 74 is the first selected village within the selected subdistrict. The conditional probability of selecting Thadijhinja (given that Maila subdistrict has already been selected) is thus $p_{2|1} = p_{2|01} = a_2 Mos_{01} / \sum_j Mos_{0j} = 2 \times 74 / 403 = 0.36725$. The overall probability of selecting this village is the product of the first stage and the conditional second stage probabilities, or $p_{01} = p_{1|01} p_{2|01} = 0.33724 \times 0.36725 = 0.12385$.

The next selection is determined by adding the fractional interval to the selection number (including the undeleted decimal portions): 201.5 + 34.8 = 236.3. Dropping the decimal portion of this next selection number to obtain 236, the next selected village is Ekrahi with cumulative total 243. The conditional probability of selection of this village is $p_{2|1} = 2 \times 61 / 403 = 0.30273$, and the overall probability of selection is $p_{01} = 0.33724 \times 0.30273 = 0.10209$. Thus, the overall probabilities of selection of these two villages are not the same. Some type of correction must be made in the last stage of selection in order to recover an epsem sample design.

The selection is completed once the required $a_1 = 2$ villages have been selected. The procedure can readily be extended to more than two selections, and it can be demonstrated that it provides the correct probabilities of selection through the "rounding" procedure of dropping the decimal portion of the selection number.

The third stage conditional selection probability, $p_{3ijk}$, where the subscript $k$ denotes third stage units within the $ij$th selected second stage unit, must now be chosen to obtain an overall epsem sample selection of households. At the third stage, some number $a_{ij}$ of third stage units (households in the example) is to be selected from the actual units that are present in the selected second stage sampling unit. There are several ways to determine the appropriate sampling rate for this third stage of sampling.

One involves determining a value of $a_{ij}$ which is the same for each of the second stage sampling units, and then calculate the rate as

$$p_{3ijk} = \frac{a_{ij}}{Mos_{ij}}.$$

For example, the design calls for a target sample of approximately the same number of households to be selected in each village, and the examination of all children ages 6 months to ten years residing in those households. If this approach is used, the overall sampling rate for households is the same for all households in the study area since

$$f = p_{ijk} = p_{1i} p_{2ij} p_{3ijk}$$

$$= \frac{a\, Mos_i}{\sum Mos_i} \cdot \frac{a_i\, Mos_{ij}}{\sum_j Mos_{ij}} \cdot \frac{a_{ij}}{Mos_{ij}}$$

$$= \frac{a}{\sum Mos_i} \cdot a_i \cdot a_{ij},$$

where $\sum_j Mos_{ij} = Mos_i$. Since $a$, $a_i$, $a_{ij}$, and $\sum Mos_i$ are all fixed by the design, the sampling rate $f$ is the same for all third stage units.

In the example, the overall sampling rate was fixed at $f$ = 1 / 28, while $a$ = 5, $a_i$ = 2, and $\sum Mos_i$ = 5,975 . Hence, solving for $a_{ij}$,

$$a_{ij} = \frac{f \sum Mos_i}{a\, a_i}$$

$$= \frac{(1/28) \times 5{,}975}{5 \times 2}$$

$$= 21.339.$$

Within each village, the sampling rate for households would be calculated separately using the fixed $a_{ij}$ = 21.339. For example, in Thadijhinja village, the within village sampling rate would be 21.339 / 74 = 1 / 3.4678, and a systematic selection with fractional interval 3.4678 could be used to select sample households from a list of all households prepared for the village at the time of the visit for examinations. In Ekrahi village, the within rate would be 21.339 / 61 = 1 / 2.8586.

Alternatively, the within secondary sampling rate can be calculated directly from the fixed overall sampling rate and the conditional probabilities of selecting first and second stage sampling units. In particular,

$$p_{3ijk} = \frac{f}{p_{1i} p_{2ij}}.$$

For Thadijhinja village the conditional probability within the village is $p_{3ijk}$ = (1 / 28) / (0.33724 × 0.36725) = 1 / 3.4678, while in Ekrahi village the rate is $p_{3ijk}$ = (1 / 28) / (0.33724 × 0.30273) = 1 / 2.8586.

When these within village sampling rates are applied to lists of households from the village, the resulting samples of households will be approximately the same size across villages. If the actual number of households within each village is the same as the measure of size for

the village, $Mos_{ij}$, then the within village sample of households will be identical across villages. For example, suppose that when Thadijhinja village is visited, exactly 74 households were found. Then the sample of households would consist of 21 or 22 of those households selected at random or systematically from the list of households in the village. On the other hand, suppose that 90 households were actually found. Then the sample would consist of 90 × (1/3.4678) = 25.953, either 25 or 26 households selected from the household list. If fewer than 74 households were found, then applying the within sampling rate would result in fewer than 21 or 22 households being selected. Since the villages will increase or decrease in size at different rates, there will be variation in sample sizes across villages. However, it is expected on average that the differences from village to village will be small and produce only minor fluctuations in workload.

One of the villages within the selected subdistrict may present a problem for the last stage of selection because it has a small measure of size. In Table 2, village 6, Machijhitkaiya, has a measure of size of only 21 households. If this village had been selected, the third stage conditional probability of selection needed to maintain epsem sampling would have been greater than 1.0. Since we cannot sample at a rate greater than 1.0, we could not maintain epsem for this village and would be forced into some type of weighting adjustments during analysis. To avoid this problem, we can set a minimum size for each village, say 22 households in the example. When a village is encountered that has a measure smaller than the minimum, it must be combined with one or more other villages to form a new secondary sampling unit which is larger than the minimum size. In the example, village 6 could have been combined with village 5 to create a new sampling unit with a measure of 58 + 21 = 79 households. If either village 5 or 6 were then selected, the subsampling at the last stage of selection would be from all households in both villages.

Implicit stratification has been employed in this example to the extent that the subdistricts and village lists are ordered by characteristics which might be correlated with study findings. But explicit stratification could also have been used. For example, the subdistricts might be divided into two strata and three subdistricts selected from one stratum and two from the other. PPS selection methods as described here must be adapted to the situation in which the strata are of different size (a topic beyond the scope of this discussion). Alternatively, the district itself may represent a stratum of subdistricts. A slight modification of the notation is needed to then identify the separate strata: a subscript $b$ would be needed to denote the $b$th stratum. The sampling rates may vary across these strata, although maintaining epsem across strata is likely to be preferred. One might also consider selecting a different number of subdistricts in each stratum, and thus $a_b$ would denote the number of subdistricts selected in each stratum.

In summary, a recommended procedure for selecting a three-stage epsem sample of households using PPS selection methods is as follows:

1) Compute the overall target sampling rate $f$.

2) Determine the number $a_h$ of first stage units to be selected in each stratum. Determine the number of secondary sampling units, say a′, to be selected within each first stage unit (this is likely to be the same across strata).

3) Select $a_h$ first stage units in each stratum with PPS and record the first stage selection probability $p_{1hi}$ for each selected unit.

4) Select a′ secondary sampling units in each selected first stage unit using PPS selection methods and record the second stage conditional selection probability $p_{2hij}$ for each selected unit.

5) For each selected secondary sampling unit, compute the third stage conditional probability needed to obtain an epsem sample.

6) After listing the households within each selected secondary sampling unit, apply the within selection rate or interval (*i.e.*, $1/p_{3hijk}$) to choose a sample of households.

This basic approach can be modified to handle only two or more than three stages of sampling. It can handle sampling situations in which the secondary sampling units must be created by survey staff rather than obtained from other sources. The last stage of selection may also be clustered to avoid difficulty recruiting children from one household but not the one next door. For example, rather than listing individual households, which may be difficult to do in many developing countries in which there is no housing unit numbering system, geographic areal units that are clusters of households can be identified within the village. Such clusters must have unambiguously defined borders that can be readily identified in the field, and should have approximately the same number of households each. The final stage of sampling can then be applied to these clusters, with examinations conducted with children in all of the households in the selected clusters. The clusters may be of small size, such as 4 to 10 households, or they may be much larger, such as from 25 to 200. The size of the cluster depends on the cost and error characteristics of the survey itself.

**Estimation.** With the complex sample design described here, the computation of estimates of even simple quantities such as proportions can be elaborate and difficult to implement if the complete complexity of the design is taken into account. We recommend a

simplified approach suitable for estimates from the design described here for a single type of estimator, confidence intervals for proportions.

Suppose that we are interested in estimating the proportion of children $P$ for whom Bitot's spots is reported. For each child in the sample, two variables would be created: (1) an indicator which is equal to 1 if the child had Bitot's spots and equal to 0 otherwise, and (2) a variable equal to 1 for all children in the dataset. Let $y_{hijk}$ denote the indicator for Bitot's spots for the $k$th child in the $j$th village in the $i$th subdistrict in stratum $h$, and let $x_{hijk}$ denote constant variable for the same $hijk$th child. The $y_{hijk}$ and $x_{hijk}$ values are summed across children, villages, subdistricts, and strata, and the estimated the proportion of children who reportedly have Bitot's spots is computed as the ratio of those two sums:

$$r = \frac{\sum_h \sum_i \sum_j \sum_k y_{hijk}}{\sum_h \sum_i \sum_j \sum_k x_{hijk}}$$

$$= \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}}$$

$$= \frac{y}{x}.$$

There are several features of this estimator that are important to note. For one, it is a ratio of the two sums over all children and not a sum of ratios for each child. This estimation procedure for proportions for the total population can be adapted for estimating proportions for subgroups as well. For instance, the summation in the numerator and denominator of the ratio can be made only for male or female children. Missing data for some individuals requires a modification also. If the numerator variable is missing (*e.g.*, the Bitot's spots status of a child is unknown), then the child does not contribute to the summation in the numerator or the denominator. Finally, a similar approach can be used to estimate means as well. Instead of the indicator variable in the numerator, the value of the variable for which the mean is to be computed (*e.g.*, serum vitamin A level) is substituted.

In order to assess the reliability of the estimated proportion, an estimate of sampling variability is needed. Estimating sampling variability requires that the stratification and clustering features of the sample design be considered as well. To estimate the variance of the estimated proportion $r$, estimates of the variance of the numerator and

denominator as well as of the covariance between numerator and denominator are needed. Let $y_{hi}$ and $x_{hi}$ denote the summation of the numerator and denominator variables across children and villages in the $h$th stratum and $i$th subdistrict:

$$y_{hi} = \sum_j \sum_k y_{hijk} \text{ and } x_{hi} = \sum_j \sum_k x_{hijk}.$$

Then the variances and covariance of the numerator and denominator sums in $r$ are

$$\text{var}(y) = \sum_h \left( \frac{a_h(1 - f_h)}{a_h - 1} \right) \times \left( \sum_i y^2_{hi} - \frac{\left( \sum_i y_{hi} \right)^2}{a_h} \right),$$

$$\text{var}(x) = \sum_h \left( \frac{a_h(1 - f_h)}{a_h - 1} \right) \times \left( \sum_i x^2_{hi} - \frac{\left( \sum_i x_{hi} \right)^2}{a_h} \right), \text{ and}$$

$$\text{cov}(y,x) = \sum_h \left( \frac{a_h(1 - f_h)}{a_h - 1} \right) \times \left( \sum_i y_{hi} x_{hi} - \frac{\left( \sum_i y_{hi} \right)\left( \sum_i x_{hi} \right)}{a_h} \right),$$

where $f_h$ is the sampling rate within the $h$th stratum. The factor $(1 - f_h)$ is referred to as the finite population correction, and we recommend that it be ignored unless the within stratum sampling rate for a given stratum is greater than 1 / 20 or perhaps 1 / 10.

These variances and the covariance can then be combined to compute an estimate of the *sampling variance* of the estimated proportion $r$:

$$\mathrm{var}(r) = (\mathrm{var}(y) + r^2\,\mathrm{var}(x) - 2\,r\,\mathrm{cov}(y, x))/x^2.$$

The *standard error* of the estimated proportion is computed as the square root of the sampling variance: $\mathrm{ste}(r) = \sqrt{\mathrm{var}(r)}$. The standard error is then used to form *confidence intervals* for $P$. The interval defined by

$$r - z_{(1-\alpha/2)}\,\mathrm{ste}(r) \text{ and } r + z_{(1-\alpha/2)}\,\mathrm{ste}(r),$$

where $z_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ value of the standard normal distribution, forms a $(1 - \alpha/2) \times 100$ percent confidence interval for the population proportion $P$. In particular, if a large number of samples were selected using exactly the same sample design, we expect that for $(1 - \alpha/2) \times 100$ percent of the samples the confidence interval constructed in this way will contain the true but unknown population proportion of interest.

A simpler approach to estimating the sampling variance of a proportion would be to use the standard formula based on the assumptions of ***simple random sampling***, or **srs**, in selecting the observations: $\mathrm{var}_{srs}(r) = r(1 - r)/x$. However, given the clustering in the sample design, it is likely that the simple random sampling estimate of sampling variance will be too narrow. Hence, confidence intervals formed using the simple random sampling estimates of sampling variance will be too small, and we will claim greater precision in the sample results than is actually warranted.

Sampling statisticians refer to the ratio of the sampling variance of the estimate to the sampling variance computed under simple random sampling assumptions as the ***design effect***:

$$\mathrm{deff} = \frac{\mathrm{var}(r)}{\mathrm{var}_{srs}(r)}.$$

For a multistage sample design, the design effect will be greater than 1.0, indicating that the effect of the clustered sample design has been to increase sampling variances relative to a simple random sample of the same size. If the design effect is anticipated to be substantial, an adjustment to the sample size may be needed to compensate for the larger variances expected from the clustered sample selection.

## 2. Sample Size Determination

The first step in determining sample size for a sample survey is to specify the level of reliability, or precision, needed for the estimates of proportions that are to be obtained from the survey. It is important to be aware that a sample size determined to meet specifications for the precision of an estimate of the proportion of children with Bitot's spots in a certain region or district would not be large enough to estimate with the same precision that same proportion for boys or for all children in a subdistrict. If a certain precision is required of estimates in these subgroups, then determination of sample size must be done at the subgroup level.

The sample size $n$ to be drawn from a population of size $N$ is to be determined so that we will be confident that the estimated proportion $r$ differs from the true population proportion $P$ by no more than a specified amount $d$. There is a small risk $\alpha$ that the actual difference is larger than $d$; that is, the probability that the estimated proportion differs from the true population proportion by more than the difference $d$ is $\alpha$:

$$\Pr(|r - P| \geq d) = \alpha.$$

Alternatively, we may wish to control the error relative to the size of the population proportion $P$, in which case we have

$$\Pr(|r - P| \geq \varepsilon P) = \alpha,$$

where $\varepsilon$ is the size of the relative error we are willing to tolerate. For example, if we want the estimated proportion $r$ to depart from the population proportion $P$ by no more than 5 percent of $P$, then $\varepsilon = 0.05$.

From simple random sampling theory, a first approximation to $n$ is given by

$$n = z^2 P_0(1 - P_0) / d^2,$$

or, in terms of relative error,

$$n = z^2 (1 - P_0) / \varepsilon^2 P_0,$$

where $P_0$ is the best *a priori* knowledge or guess about the unknown population proportion $P$ and $z = z_{(1-\alpha/2)}$ is the abscissa of the standard normal distribution which cuts off an area of $\alpha / 2$ in each of the tails of the distribution.

The sample size formula which specifies the difference $d$ is relatively insensitive to errors in the *a priori* value of $P_0$. If there is some uncertainty about a reasonable value for $P_0$, it is better to choose a value that is closer to 0.5. For example, suppose that the value of $P$ is suspected to be somewhere in the range 0.10 to 0.15; a value of 0.15

is more conservative (safer) for estimating sample size. On the other hand, if $P$ is thought to be in the range 0.60 to 0.70, it is better to use $P_0 = 0.60$ (*i.e.*, closer to 0.5). However, if the formula which specifies relative error $\varepsilon$ is used, it is safer to use the value closer to 0.0 ($P_0$ = 0.10 and 0.60, in this example).

When the sample is being selected from a small population, or the sample size is a major portion of the population, the formula for sample size may be adjusted to account for the finite nature of the population. In particular,

$$n' = n / (1 + n / N)$$

may be used to obtain a sample size adjusted for the finite nature of the population. This adjustment is not often needed, and we recommend that unless the original estimated sample size is more than 10 percent of the population size, this adjustment be ignored.

For confidence expressed as "reasonably certain", we often use a risk of error of $\alpha = 1 / 20 = 0.05$, which corresponds to $z = z_{(1-0.05/2)} = z_{0.975} = 1.96$. We are "virtually certain" if we use $\alpha = 3 / 1{,}000$ which corresponds to $z_{0.9985} = 3.0$.

The clustered nature of the recommended sample design may yield a substantial design effect requiring an adjustment which increases the sample size determined from simple random sampling theory. Results from previous surveys of a similar design for the same characteristic (*e.g.*, Bitot's spots), or from a pilot study, may be available to estimate the design effect. In the absence of such results, assume a design effect of at least *deff* = 1.5, and use $n \times$ *deff* as the best estimate of the required sample size.

In Tables 3 and 4 of this section values are given of sample size $n$ for selected values of $P_0$, $\alpha$, and $\varepsilon$; selected $z$-values from the normal distribution for various $\alpha$; and a summary of the steps required to use the tables including the finite population adjustment and adjustment for design effect.

**Example 1.** Suppose that before beginning a vitamin A intervention program in a region with 10,000 preschool children, we wish to estimate the prevalence $P$ of preschool children in the region with Bitot's spots. Available information suggests that the prevalence is about 3 percent (*i.e.*, $P_0 = 0.03$).

**a.** We want to be "reasonably certain" that our survey provides an estimate $r$ such that $|r - P| < 0.01$; that is, the interval $r \pm 0.01$ will include $P$ except for a 1/20 chance. Here, $P_0 = 0.03$, $d = 0.01$, $\alpha = 0.05$, $z = 1.96$, and $N = 10{,}000$. Hence,

$$n = (1.96)^2(0.03)(0.97) / (0.01)^2 = 1{,}118,$$

and, adjusting for the finite population,

$$n' = 1{,}118 / (1 + 1{,}118 / 10{,}000) = 1{,}006.$$

## Appendix Table 3

**Sample size $n$ based on $P_0$, $\alpha$, $\varepsilon$, and simple random sampling, required to estimate a proportion P.**

| | | $P_0$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\varepsilon$ | 0.5 | 0.25 | 0.15 | 0.05 | 0.025 | 0.005 | 0.0001 |
| 0.050 | 0.50 | 16 | 47 | 88 | 295 | 600 | 3060 | 153700 |
| | 0.25 | 62 | 185 | 350 | 1170 | 2400 | 12240 | |
| | 0.15 | 170 | 515 | 970 | 3250 | 6660 | 34000 | |
| 0.010 | 0.50 | 27 | 80 | 150 | 505 | 1040 | 5300 | 266300 |
| | 0.25 | 110 | 320 | 605 | 2030 | 4160 | 21200 | |
| | 0.15 | 295 | 890 | 1680 | 5620 | 11540 | 59000 | |
| 0.003 | 0.50 | 37 | 110 | 205 | 685 | 1410 | 7170 | 360000 |
| | 0.25 | 145 | 435 | 820 | 2740 | 5620 | 28660 | |
| | 0.15 | 400 | 1200 | 2270 | 7600 | 15600 | 79600 | |

Step 1 Determine $N$, the size of the population to be sampled, $P_0$, the best prior estimate of the true unknown population proportion P, $\alpha$, the risk that the estimated proportion differs from P by $\varepsilon P_0$, and $\varepsilon$, the maximum acceptable relative error.

Step 2 For the selected $\{P_0, \alpha, \varepsilon\}$, determine $n$ from the table for simple random sampling

Step 3 Apply the finite population adjustment, $n' = n / (1 + n/N)$

Step 4 Apply deff, the best prior estimate of the design effect (or, assume deff = 1.5 if no estimate is available), to get the final estimate of the required sample size: Sample size = $n' \times$ deff

No information is available on *deff*, so we arbitrarily adjust $n'$ upward by a factor of 1.5 to obtain a final sample size estimate 1,006 × 1.5 = 1,510. It is clear that the accuracy of the initial estimate based on simple random sampling is not critical, since the inflation factor to adjust for effects of clustering will often be arbitrary.

**b.** If instead we wish to be virtually certain of our precision (*i.e.*, $\alpha = 3 / 1{,}000$ and $z = 3.0$), then

$$n = (3.0)^2(0.03)(0.97) / (0.01)^2 = 2{,}619,$$

and, adjusting for the finite population,

$$n' = 2{,}619 / (1 + 2{,}619 / 10{,}000) = 2{,}075.$$

Applying the design effect factor, the final sample size estimate is 2,075 × 1.5 = 3,113.

**c.** If uncertainty about the existing prevalence suggests that $P$ may be possibly as large as 5 percent, then the estimates in (a.) and (b.) become 2,315 and 4,493, respectively.

**Example 2.** Suppose that we have completed the survey

## Appendix Table 4

**Selected z-values from the standard normal distribution, useful in estimation of sample size.**

| $\alpha$ | $z_{(1-\alpha)}$ | $z_{(1-\alpha/2)}$ |
|---|---|---|
| 0.2 | 0.84 | 1.28 |
| 0.1 | 1.28 | 1.64 |
| 0.05 | 1.64 | 1.96 |
| 0.01 | 2.33 | 2.58 |
| 0.003 | 2.75 | 3.0 |
| 0.001 | 3.1 | 3.3 |

planned in Example 1 which, using the sample design of Section 1, yields the estimates $r_1 = 0.025$, $ste(r_1) = 0.005$, and $deff = 1.2$. Vitamin A intervention is subsequently done, and another survey is planned to evaluate the intervention. We want to be reasonably certain that the post–intervention survey provides an estimate $r_2$ which, if it is 0.01 or smaller, will be found significantly different from $r_1$ in a one–sided test of statistical significance.

In planning a statistical test of the hypothesis that the difference $r_1 - r_2$ is zero, against the alternative hypothesis that $r_1 > r_2$, we can control the risks of making two errors: $\alpha$, the risk of incorrectly deciding the difference is not zero when it really is; and $\beta$, the risk of incorrectly deciding the difference is zero when it really is not. Both $\alpha$ and $\beta$ are to be small.

The test statistic for the test of a difference between $r_1$ and $r_2$ is

$$Z = \frac{(r_1 - r_2)}{\sqrt{ste^2(r_1) + ste^2(r_2)}}.$$

When $Z$ is greater than $z_{(1-\alpha)} = z_{0.95} = 1.645$, we conclude with reasonable certainty that $r_1$ is greater than $r_2$. We want a sample size for the post–intervention survey which, will provide a statistically significant result (i.e., $Z > 1.645$) with probability $0.8 (= 1 - \beta)$ if $r_2 = 0.01$. Now,

$$ste^2(r_2) = r_2(1 - r_2) \times deff / n_2,$$

where $n_2$ is the sample size of the post–intervention survey. Given that we have estimated $r_1$, $ste(r_1)$, and $deff$, and that we have chosen $r_2 = 0.01$, $\alpha = 0.05$ ($z_{(1-\alpha)} = 1.645$), and $\beta = 0.2$ ($z_{(1-\beta)} = 0.842$), we can use a result from large–sample hypothesis testing theory, $z_{(1-\beta)} + Z = z_{\alpha}$, to solve for the value of $n_2$:

$$n_2 = \frac{r_2(1 - r_2) \times deff}{\{(r_1 - r_2)/(z_{(1-\alpha)} + z_{(1-\beta)})\}^2 - ste^2(r_1)}$$

$$= \frac{(0.01)(0.99)(1.2)}{(0.015/(1.645 + 0.842))^2 - (0.005)^2}$$

$$= 1044.$$

A sample of this size will yield $ste(r_2) = 0.0034$, and thus, the 95 percent confidence interval for $P$ at the post–intervention time is $r_2 \pm (1.96)ste(r_2) = 0.01 \pm 0.007$. Even though $n_2$ is adequate for the test of significance, we may wish to increase $n_2$ to improve the precision of

# Recommended Reading


Arroyave, G.; Aguilar, J.R.; Flores, M.; Guzman, M.A. Evaluation of sugar fortification with vitamin A at the national level. Pan American Health Organization Scientific Publication No. 384, 1979. Washington, D.C.

Arroyave, G.; Bauernfeind, J.C.; Darby, W.J.; Karyadi, D.; Latham, M.C.; McLaren, D.S.; Olson, J.A.; Pettis, S.T.; Pirie, A.; Sommer, A.; Teply, L.J.; Underwood, B.A. Guidelines for the eradication of vitamin A deficiency and xerophthalmia. Report of the International Vitamin A Consultative Group (IVACG). Washington, D.C.: The Nutrition Foundation, Inc.; 1977.

Arroyave, G; Chichester, C.O.; Flores, H.; Glover, J.; Mejia, L.A.; Olson, J.A.; Simpson, K.L.; Underwood, B.A. Biochemical methodology for the assessment of vitamin A status. Report of the International Vitamin A Consultative Group (IVACG). Washington, D.C.: The Nutrition Foundation, Inc.; 1982.

Control of vitamin A deficiency and xerophthalmia. Report of a Joint WHO/UNICEF/USAID/Helen Keller International/IVACG Meeting. WHO Tech. Rep. Ser. 672; 1982.

Cowan, B.; Dhanoa, J. The prevention of toddler malnutrition by home-based nutrition, health education. In: McLaren, D.S., ed. Nutrition in the community. Chichester, England: John Wiley and Sons; 1983: p. 339.

Data needs for assessing the nutritional effects of agricultural and rural development projects: A paper for project planners. Nutrition in Agriculture, No. 4, Rome: FAO; 1983.

Hornick, R.C. Nutrition education: state-of-the-art review. Administrative Committee on Coordination-Subcommittee on Nutrition (ACC-SCN). Rome: FAO Food Policy and Nutrition Division; 1985.

Israel, R.C.; Tighe, J.P.N. Nutrition education: state-of-the-art, a review and analysis of the literature. Paris: UNESCO Nutrition Education Series Issue 7; 1984.

Jones, E.M.; Munger, S.J.; Altman, J.W. A field guide for evaluation of nutrition education. Washington, D.C.: Office of Nutrition, Technical Assistance Bureau, Agency for International Development; 1975.

Klein, R.E., et al. editors. Evaluating the impact of nutrition and health programs. Proceedings of an international conference sponsored by the Pan American Health Organization; 1979; Isla Cintadora, Panama. 462 p. Available from: Plenum Press, New York.

Latham, M.C.; Solon, F.S. Vitamin A deficiency control in the Philippines. In: Bauernfeind, J.C., ed. Vitamin A deficiency and its control. Orlando, Florida: Academic Press, Inc.; 1986.

McLaren, D.S.; Ballantine, E.J.; ten Doesschate, J.; Hodges, R.E.; Pararajasegaram, R.; Sommer, A.; Venkataswamy, G. The symptoms and signs of vitamin A deficiency and their relationship to applied nutrition. Report of the International Vitamin A Consultative Group (IVACG). Washington, D.C.: The Nutrition Foundation, Inc.; 1981.

Rossi, P.H.; Freeman, H.E.; Wright, S.R. Evaluation. A systematic approach. London: Sage Publications, Inc.; 1979.

Sahn, D. Methods for evaluating the nutritional impact of food aid projects: lessons from past experience. International Food Policy Research Institute, Washington, D.C. Food and Nutrition Bulletin Volume 6 No. 3 September, 1984, pages 1–16. 1984.

Schurch, B., editor. Evaluation of nutrition education in third world communities. Nestle Foundation Publication Series. Bern: Hans Huber Verlag; 1983.

Selecting interventions for nutritional improvement: A manual. Nutrition in Agriculture No. 3, Rome, FAO; 1983.

Vitamin A deficiency and xerophthalmia. Report of a Joint WHO/USAID Meeting. WHO Tech. Rep. Ser. 590; 1976.

West, K.P.; Sommer, A. Periodic large oral doses of vitamin A for the prevention of vitamin A deficiency and xerophthalmia: a summary of experiences. Report of the International Vitamin A Consultative Group (IVACG). Washington, D.C.: The Nutrition Foundation, Inc.; 1984.

Zeitlin, M. Directions for the evaluation of nutrition education. In: Shack, K.W., ed. Teaching nutrition in developing countries. Santa Monica, California: Meals for Millions Foundation, pages 169-193. 1977.

The following monographs are published by the International Vitamin A Consultative Group:

- *Guidelines for the Eradication of Vitamin A Deficiency and Xerophthalmia* (1977)
- *Recent Advances in the Metabolism and Function of Vitamin A and Their Relationship to Applied Nutrition* (1979)
- *The Safe Use of Vitamin A* (1980)
- *The Symptoms and Signs of Vitamin A Deficiency and Their Relationship to Applied Nutrition* (1981)
- *Biochemical Methodology for the Assessment of Vitamin A Status* (1982)
- *Reprints of Selected Methods for the Analysis of Vitamin A and Carotenoids in Nutrition Surveys* (1982)
- *Periodic Large Oral Doses of Vitamin A for the Prevention of Vitamin A Deficiency and Xerophthalmia: A Summary of Experiences* (1984)
- *The Safe Use of Vitamin A by Women During the Reproductive Years* (1986)
- *Biochemical Methodology for the Assessment of Carotenes* (1987)
- *A Decade of Achievement: The International Vitamin A Consultative Group (IVACG) 1975–1985* (1987)
- *Guidelines for the Use of Vitamin A in Emergency and Relief Operations* (1988)
- *Vitamin A Supplements: A Guide to Their Use in the Treatment and Prevention of Vitamin A Deficiency and Xerophthalmia* (published by the World Health Organization in conjunction with IVACG and UNICEF, 1988)
- *Guidelines for the Development of a Simplified Dietary Assessment to Identify Groups at Risk for Inadequate Intake of Vitamin A* (1989)

These reports are available free of charge to developing countries and for $3.50 (U.S.) to developed countries. Copies can be ordered from the IVACG Secretariat:

IVACG Secretariat
International Life Sciences Institute–Nutrition Foundation
1126 16th Street, N.W.
Washington, D.C. 20036
USA